DOCUMENT RESUME

ED 452 297 UD 034 118

AUTHOR	Krueger, Alan B.; Hanushek, Eric A.
TITLE	The Class Size Policy Debate. Working Paper No. 121.
INSTITUTION	Economic Policy Inst., Washington, DC.
PUB DATE	2000-10-00
NOTE	50p.; Introduction by Richard Rothstein.
AVAILABLE FROM	Economic Policy Institute, 1660 L Street, N.W., Suite 1200, Washington, DC 20036 ($10). Web site: http://www.epinet.org.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Class Size; Cost Effectiveness; Econometrics; *Educational Economics; *Educational Research; Elementary Secondary Education; Expenditure per Student; Public Education
IDENTIFIERS	Student Teacher Achievement Ratio Project TN

ABSTRACT

These papers examine research on the impact of class size on student achievement. After an "Introduction," (Richard Rothstein), Part 1, "Understanding the Magnitude and Effect of Class Size on Student Achievement" (Alan B. Krueger), presents a reanalysis of Hanushek's 1997 literature review, criticizing Hanushek's vote-counting method and suggesting it would be better to count each publication as a single study rather than counting separately each estimate within a publication. This method of analyzing the research suggests a strong connection between school resources used to reduce class size and student outcomes. It also discusses the effects of expenditures per student and looks at economic criterion (Lazear's theory of class size, benefits and costs of educational resources, the critical effect size, and caveats). Part 2, "Evidence, Politics, and the Class Size Debate" (Eric A. Hanushek), focuses on: the history of class size reduction; econometric evidence; the Tennessee Class Size Experiment (Project STAR, or the Student/Teacher Achievement Ratio Study); and policy calculations. The author claims that his own method of counting each estimate as a separate study is the most valid method and that effects of class size reduction will be small and expensive. An appendix discusses the econometric data. (Papers contain references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 452 297

WORKING PAPER

THE CLASS SIZE POLICY DEBATE

Understanding the magnitude and effect of class size on student achievement
by Alan B. Krueger

Evidence, politics, and the class size debate
by Eric A. Hanushek

Introduction by Richard Rothstein

Working Paper No. 121 • October 2000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY:

P. Watson
Economic Policy Institute

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

UD034118

Economic Policy Institute

BEST COPY AVAILABLE

Alan B. Krueger
Princeton University and National Bureau of Economic Research

Prof. Krueger's paper is a revised and extended version of a paper that was originally prepared for a conference sponsored by Temple University's Center for Research in Human Development and Education titled, "What Do We Know About How to Make Small Classes Work?" held December 6-7, 1999 in Washington, D.C. The paper was written while Prof. Krueger was on leave at the Center for Advanced Study in the Behavioral Sciences at Stanford University. He is grateful to Diane Whitmore and Michael Watts for excellent research assistance, to Victor Fuchs for helpful comments, and to Eric Hanushek for providing the data used in Section I. Jesse Rothstein provided valuable editorial assistance. The Center for Advanced Study in the Behavioral Sciences, Temple's Center for Research in Human Development and Education, and Princeton's Industrial Relations Section provided financial support. The author bears sole responsibility for the views expressed in the paper and for any errors.

Eric A. Hanushek
Stanford University, University of Texas at Dallas, and National Bureau of Economic Research

Helpful comments on Prof. Hanushek's paper were provided by John Kain, Steve Landsburg, Ed Lazear, Terry Moe, Paul Peterson, Macke Raymond, and Steve Rivkin.

Richard Rothstein
Research associate, Economic Policy Institute; adjunct professor of public policy, Occidental College in Los Angeles; senior correspondent, The American Prospect; *weekly education columnist,* The New York Times

Copyright © 2000 Economic Policy Institute

1660 L Street NW Suite 1200

Washington, DC 20036

www.epinet.org

TABLE OF CONTENTS

INTRODUCTION *by Richard Rothstein* 1

PART I: UNDERSTANDING THE MAGNITUDE AND EFFECT OF CLASS SIZE ON STUDENT ACHIEVEMENT *by Alan B. Krueger* 5

I. Reanalysis of Hanushek's literature review 7

Expenditures per student 13

Summing up 14

II. Economic criterion 15

Lazear's theory of class size 15

Benefits and costs of educational resources 16

The 'critical effect size' 20

Caveats 20

III. Conclusion 22

Endnotes 23

References 25

PART II: EVIDENCE, POLITICS, AND THE CLASS SIZE DEBATE *by Eric A. Hanushek* 27

I. The history of class size reduction 28

II. Econometric evidence 30

III. The Tennessee Class Size Experiment (Project STAR) 38

IV. Policy calculations 40

V. Conclusions 42

Appendix: Issues with the econometric data 43

Endnotes 43

References 45

INTRODUCTION

Richard Rothstein

For three decades, a belief that public education is wasteful and inefficient has played an important role in debates about its reform. Those who have proposed new spending programs for schools to improve student achievement have been on the defensive. The presumption has been that changes in structure and governance of schools — like choice, vouchers, charter schools, standards, accountability, and assessment — are the only way to improve student outcomes. Traditional interventions, like smaller class size and higher teacher salaries, have been presumed ineffective.

Voters and state and local political leaders have never been as impressed with this statement of alternatives as have national policy makers and scholars. Throughout the last third of the 20th century, when the idea that "money makes no difference" held sway in academic circles, spending in public education increased at a steady rate, and class sizes declined. But, as we showed in a 1995 Economic Policy Institute report, *Where's the Money Gone?*, the spending has increased more slowly than most people believe. It can't be known whether the rate would have been more rapid in the absence of an academic consensus regarding public education's inefficiency.

The leading proponent of the prevailing view that money doesn't make a difference has been Eric A. Hanushek, now of the Hoover Institution. Dr. Hanushek has played two roles. As a scholar, he has conducted a series of influential literature reviews that support the conclusion that increased spending in general, and smaller class size in particular, do not "systematically" lead to improved student achievement. There have been hundreds of research studies that attempt to assess the relationship of spending and achievement. Dr. Hanushek has found that, in some cases, the relationship is positive, but in others no positive relationship can be discerned, either because the relationship is negative or because it is statistically insignificant.

These findings have led Dr. Hanushek to play another role — as a very visible public advocate for restraining the growth of spending in public schools. He chaired a task force of the Brookings Institution, leading to the publication of *Making Schools Work: Improving Performance and Controlling Costs*, a very influential 1993 book that asserts, "Despite ever rising school budgets, student performance has stagnated....[I]n recent years the costs of education have been growing far more quickly than the benefits." Dr. Hanushek has testified in many state court cases regarding the equity and adequacy of school spending, generally in support of the proposition that increased funds are not a likely source of improved student achievement. He is also frequently cited in newspapers and magazines in support of this proposition.

Dr. Hanushek's academic research, inventorying and summarizing existing studies of the relationship between spending and achievement, does not inexorably lead to conclusions about the desirability of restraining school spending. Even if his conclusion about the lack of a "systematic" relationship is unchallenged, it remains the case that some studies show a positive relationship, and therefore it might be possible to determine when, and under what conditions, higher spending produces student achievement. Dr. Hanushek states as much in almost all of his academic publications, but with the caveat that "simply knowing that some districts might use resources effectively does not provide any guide to effective policy, unless many more details can be supplied." However, Dr. Hanushek's research has not led a generation of scholars and policy makers to seek to supply these details. Rather, the impact has mostly been to encourage policy makers to look away from resource solutions and toward structural and governance changes.

In recent years, the most important challenge to this dominant trend has arisen because of an unusual

experiment (STAR, or the Student/Teacher Achievement Ratio study) conducted by the state of Tennessee. Attempting to determine whether achievement would increase with smaller class sizes, the state legislature authorized schools to volunteer to participate in an experiment whereby they would receive additional funds for lower class sizes for kindergarten to third grade, provided that students and teachers were randomly assigned to regular (large) or small classes.

The result was significantly enhanced achievement for children, especially minority children, in smaller classes. This single study persuaded many scholars and policy makers that smaller classes do make a difference, because the study was believed to be of so much higher quality than the hundreds of non-experimental studies about which Dr. Hanushek had relied for his summaries. Most theoreticians have long believed that conducting true randomized field experiments is the only valid method for resolving disputes of this kind. The reason is that, in non-experimental studies, comparisons between groups must ultimately rely on researchers' assumptions about similarity of the groups' characteristics. This makes the studies subject to errors from mis-specification (for example, assuming that black students who receive free or reduced-price lunch subsidies are similar in relevant respects to white students who receive these subsidies) or from omitted variables (for example, failing to recognize that parental education levels are important determinants of student achievement).

Randomized field trials, on the other hand, avoid these flaws because, if treatment and control groups are randomly selected from large enough populations, researchers can assume that their relevant characteristics (whatever those characteristics may be) will be equally distributed between the two groups. In a non-experimental study, retrospective comparison of student achievement in small and large classes may lead to the conclusion that small classes are superior only because of some unobserved characteristic that distinguishes the two groups, besides the size of their classes. In an experimental study, results are more reliable because the unobserved characteristics, whatever they may be, are evenly distributed.

It is hard to avoid the conclusion that, however valid the Tennessee study will ultimately be judged to have been, enthusiasm for it has been somewhat excessive because another principle of scientific experimentation has been largely ignored: results should be confirmed over and over again before acceptance, in different laboratories where unobserved laboratory conditions may be different. In this case, even if the Tennessee results are entirely reliable, policy conclusions are being drawn that go beyond what the Tennessee results can support. For example, the Tennessee study showed that small classes are superior to large ones, but because both types of classes were mostly taught by teachers trained in Tennessee colleges, earning similar salaries on average, it is possible that the results would not be reproduced by teachers trained in different institutions, having different qualifications, or earning higher or lower salaries. As another example, the Tennessee study found that student achievement was higher in classes of about 16 than in classes of about 24. The Tennessee study itself cannot suggest whether other degrees of reductions in class size would also boost achievement.

Nonetheless, the Tennessee study has had great influence on policy makers. In California, the governor and legislature made the needed additional money available to all schools that reduced class sizes to 20 in grades K-3. California previously had nearly the largest class sizes in the nation, so the reductions were substantial. But implementation of this policy illustrates the dangers of rushing to make policy changes based on limited research. Because California increased its demand for elementary school teachers so suddenly, many teachers without training or credentials were hired. At the same time, many experienced teachers, working in lower-income and minority communities, transferred to districts with more affluent and easier-to-teach students, taking advantage of the vast numbers of sudden openings in suburban districts. Class size reduction therefore had the result in California of reducing the average experience (and, presumably quality)

of K-3 teachers in the inner city. Nonetheless, since the implementation of the class size reduction policy, test scores in California schools, including schools that are heavily minority and low income, rose. But because California simultaneously implemented other policy changes (abolition of bilingual education, a stronger accountability system), it is uncertain to what extent class size reduction has been responsible for the test score gains.

Thus, as we enter a new decade, these two controversial lines of research — Dr. Hanushek's conclusion that there is no systematic relationship between resources and achievement, and the STAR results that smaller class sizes do make a difference — while not entirely inconsistent, are contending for public influence. In the following pages, the Economic Policy Institute presents a new critique of Dr. Hanushek's methodology by Alan Krueger, a professor of economics at Princeton, and a reply by Dr. Hanushek.

Dr. Krueger's paper has two parts. First, he criticizes Dr. Hanushek's "vote counting" method, or how Dr. Hanushek adds together previous studies that find a positive relationship and those that find none. In particular, Dr. Krueger notes that many of the published studies on which Dr. Hanushek's conclusions rely contain multiple estimates of the relationship between resources and achievement, and in particular between pupil-teacher ratio and achievement. In these cases, Dr. Hanushek counted each estimate separately to arrive at the overall total of studies that suggested either a positive, negative, or statistically insignificant effect for resources. But Dr. Krueger suggests that it would be more appropriate to count each publication as a single "study," rather than counting separately each estimate within a publication. By counting each publication as only one result, Dr. Krueger concludes that the effect of resources on achievement is much more positive than Dr. Hanushek found.

In the second part of his paper, Dr. Krueger applies the findings of the Tennessee STAR experiment to his own previous research on the effect of school spending on the subsequent earnings of adults, and to similar research conducted with British data. From assumptions about future interest rates, Dr. Krueger estimates the long-term economic benefits in greater income from class size reduction, and concludes that, with plausible assumptions, the benefits can be substantial, exceeding the costs.

In this respect, Dr. Krueger's paper is an important advance in debates about education productivity. By comparing the long-term economic benefits and costs of a specific intervention, he has shown that education policy making can go beyond an attempt to evaluate school input policies solely by short-term test score effects. While, in this preliminary exploration, Dr. Krueger has had to make substantial assumptions about the organization and financial structures of schools (assumptions he notes in "caveats" in the paper), he has defined a framework for the cost-benefit analysis of school spending for other researchers to explore, elaborate, and correct.

Dr. Hanushek responds to each of the Krueger analyses. With regard to the claim that "vote counting" should be based on only one "vote" per published study, Dr. Hanushek challenges the statistical assumptions behind Dr. Krueger's view and concludes, again, that his own method, of counting each estimate as a separate study, is more valid. Dr. Krueger's method, he suspects, was designed mainly for the purpose of getting a more positive result.

With respect to Dr. Krueger's estimates of the long-term economic effects of class size reduction, Dr. Hanushek notes that the estimates ultimately rely solely on evidence of labor market experiences of young Britons in the 1980s. "While it may be academically interesting to see if there is any plausibility to the kinds of class size policies being discussed, one would clearly not want to commit the billions of dollars implied by the policies on the basis of these back-of-the-envelope calculations."

It is unfortunate that the subject of public education has become so polarized that policy debates,

allegedly based on scholarly research, have become more contentious than the research itself seems to require. A careful reading of the papers that follow cannot fail to lead readers to the conclusion that there is substantial agreement between these antagonists. It is perhaps best expressed by Dr. Hanushek when he states,

> Surely class size reductions are beneficial in specific circumstances — for specific groups of students, subject matters, and teachers....Second, class size reductions necessarily involve hiring more teachers, and teacher quality is much more important than class size in affecting student outcomes. Third, class size reduction is very expensive, and little or no consideration is given to alternative and more productive uses of those resources.

Similarly, in his paper, Dr. Krueger states,

> The effect sizes found in the STAR experiment and much of the literature are greater for minority and disadvantaged students than for other students. Although the critical effect size differs across groups with different average earnings, economic considerations suggest that resources would be optimally allocated if they were targeted toward those who benefit the most from smaller classes.

It is difficult to imagine that Dr. Krueger would disagree with Dr. Hanushek's statement, or that Dr. Hanushek would disagree with Dr. Krueger's.

Too often, scholarship in education debates is converted into simplified and dangerous soundbites. Sometimes liberals, particularly in state-level controversies about the level, equity, or adequacy of per-pupil spending, seem to permit themselves to be interpreted as claiming that simply giving more money to public schools, without any consideration to how that money will be spent, is a proven effective strategy. In contrast, conservatives sometimes permit themselves to be interpreted as claiming that money makes no difference whatsoever, and that schools with relatively few resources can improve sufficiently simply by being held accountable for results.

But surely the debate should not be so polarized. All should be able to agree that some schools have spent their funds effectively, and others have not. All should be able to agree that targeting the expenditure of new funds in ways that have proven to be effective is far preferable to "throwing money at schools" without regard to how it will be spent. All should be able to agree that there is strong reason to suspect that minority and disadvantaged children can benefit more than others from a combination of smaller class sizes and more effective teachers. And all should be able to agree that much more research is needed to understand precisely what the most effective expenditures on schools and other social institutions might be if improving student achievement and narrowing the gap in achievement between advantaged and disadvantaged children are the goals.

It is difficult to avoid the conclusion that continued debates about whether money in the abstract makes a difference in education, without specifying how it might be spent, are unproductive. Equally true, denying that specific resource enhancements, alongside policy changes, can be an essential part of any reform agenda is also unproductive. Hopefully, the Krueger-Hanushek dialogue that follows can help to focus future debates on where spending is more effective. And it can add a new dimension to these debates, by proposing a comparison of the longer-term economic benefits of school spending, compared to its costs, that has barely begun to be explored.

PART I: UNDERSTANDING THE MAGNITUDE AND EFFECT OF CLASS SIZE ON STUDENT ACHIEVEMENT

Alan B. Krueger

At heart, questions concerning the desirability of spending more money to reduce class size involve economics, the study of how scarce resources are allocated to produce goods and services to satisfy society's competing desires. Aside from the opportunity cost of students' time, teachers are the most important, and most costly, factor of production in education. The "education production function" — that is, the relationship between schooling inputs, such as teachers per student, and schooling outputs, such as student achievement — is a special case of production functions more generally. As in other service industries, output in the education sector is hard to measure. In practice, educational output is most commonly measured by student performance on standardized tests, which is an incomplete measure for many reasons, not least because test scores are only weakly related to students' subsequent economic outcomes. Nonetheless, the output of the education sector is particularly important for the economy as a whole because as much as 70% of national income can be attributed to "human capital."[1] The education production function is thus central to understanding the economy, just as economics is central to understanding the education production function.

In recent years, a number of researchers and commentators have argued that the education production function is broken. Most prominently, in a series of influential literature summaries, Eric Hanushek (1986, 1989, 1996a, 1996b, 1997, 1998) concludes that, "There is no strong or consistent relationship between school inputs and student performance."[2] Although Hanushek never defines his criterion for a strong or consistent relationship, he apparently draws this conclusion from his findings that "studies" are almost equally likely to find negative effects of small class sizes on achievement as they are to find positive effects, and that a majority of the estimates in the literature are statistically insignificant.[3] A number of other authors have consequently concluded that the presumed failure of the education system to convert inputs into measurable outputs is an indication that incentives in public education are incapable of producing desired results. For example, John Chubb and Terry Moe (1990) argue that the "existing [educational] institutions cannot solve the problem, because they are the problem." And Chester Finn (1991) writes, "If you were setting out to devise an organization in which nobody was in command and in which, therefore, no one could easily be held accountable for results, you would come up with a structure much like American public education." In short, these critics argue that bureaucracy, unions, and perverse incentives cause public education to squander resources, severing the link between school inputs and outputs. Many observers have concluded from these arguments that it would be wasteful to put additional resources into the current public education system — either to make the system more equitable or to increase resources for all students — because they would have no effect on educational outcomes.

Hanushek's literature reviews have had widespread influence on the allocation of school resources. He has testified about his literature summaries in school financing cases in Alabama, California, Missouri, New Hampshire, New York, Maryland, New Jersey, and Tennessee, and in several congressional hearings, and his tabulations summarizing the literature have been widely cited by expert witnesses in other venues. Moreover, the presumed absence of a relationship between resources and student outcomes for the average school district has led many to support a switch to school vouchers, or a system that penalizes schools with low-achieving students.

However, a reanalysis of Hanushek's literature reviews, detailed in Section I below, shows that his

results depend crucially on the peculiar way in which he combines the many studies in the literature. Specifically, Hanushek places more weight on studies from which he extracted more estimates.

Hanushek's (1997) latest published summary of the literature on class size is based on 277 estimates drawn from 59 studies. Considerably more estimates were extracted from some studies than from others. Although the distinction between estimates and studies is often blurred, Hanushek's analysis applies equal weight to every estimate, and therefore assigns much more weight to some studies than others.[4] Hanushek's pessimistic conclusion about the performance of the education production function results in part from the fact that he inadvertently places disproportionate weight on studies that are based on smaller samples. This pattern arises because Hanushek used a selection rule that would take more estimates from studies that analyzed subsamples of a larger dataset than from studies that used the full sample of the larger dataset.

For example, if one study analyzed a pooled sample of third through sixth graders, it would generate a single estimate, whereas if another study using *the same data* analyzed separate subsamples of third graders, fourth graders, fifth graders, and sixth graders, that study would generate four estimates. Moreover, if the second study estimated separate models for black, white, and Hispanic students it would yield 12 estimates by Hanushek's selection rule. And if the study further estimated separate regressions for math and reading scores for each subsample, as opposed to the average test score, it would yield 24 estimates. As a consequence of this selection rule, the lion's share of Hanushek's 277 estimates were extracted from a small minority of the 59 studies. Specifically, 44% of the estimates come from a mere 15% of the studies. Many of these estimates are based on small subsamples of larger datasets, and are therefore very imprecise.[5] Other things being equal, estimates based on smaller samples are likely to yield weaker and less systematic results. Thus, in the example above, the 24 estimates from the second study would be considerably less precise, and therefore less likely to be statistically significant, than the single estimate from the first study; nevertheless, in Hanushek's weighting scheme the second study is given an effective weight 24 times as large as the first study.

When the various studies in Hanushek's sample are accorded equal weight, *class size is systematically related to student performance*, even using Hanushek's classification of the estimates — which in some cases appears to be problematic.

A more general point raised by the reanalysis of Hanushek's literature summary is that not all estimates are created equal. One should take more seriously those estimates that use larger samples, better data, and appropriate statistical techniques to identify the effects of class size reduction. Hedges, Laine, and Greenwald (1994) and other formal meta-analyses of class size effects reach a different conclusion than Hanushek largely because they combine estimates across studies in a way that takes account of the estimates' precision. Although their approach avoids the statistical pitfalls generated by Hanushek's method, it will still yield uninformative results if the equations underlying the studies in the literature are mis-specified. Research is not democratic. In any field, one good study can be worth more than the rest of the literature. There is no substitute for understanding the specifications underlying the literature and conducting well-designed experiments.

The largest and best designed experiment in the class size literature is Tennessee's Project STAR (Student/Teacher Achievement Ratio). According to the Harvard statistician Frederick Mosteller (1995), Project STAR "is one of the most important educational investigations ever carried out and illustrates the kind and magnitude of research needed in the field of education to strengthen schools." Studies based on the STAR experiment find that class size has a significant effect on test scores: reducing class size from 22 to 15 in the early primary grades seems to increase both math and reading test scores by about 0.2 standard deviations (see, e.g., Finn and Achilles 1990 or Krueger 1999b). In my opinion, the careful design of the STAR experiment makes these results more persuasive than the rest of the literature on class size.

Section II below considers the economic implications of the magnitude of the relationship between class size and student performance. Reducing class sizes is expensive, and it is reasonable to ask whether the benefits justify the cost. Most of the literature on class size reduction tests whether one can statistically reject the hypothesis of zero effect on performance. But for most purposes a zero effect is not a meaningful null hypothesis to test. A more appropriate question is, "How big an improvement in student performance is necessary to justify the cost?" This question is tackled here, and a provisional answer to it is then compared to the benefits from smaller classes found by the STAR experiment. The calculations described in Section II, subject to the many caveats listed there, suggest that the economic benefits of further reductions in class size in grades K-3 are at least equal to the costs.

While it is possible that a change in incentives and enhanced competition among schools could improve the efficiency of public schools, such a conclusion should rest on direct evidence that private schools are more efficacious than public schools, or on evidence that competition improves performance, not on a presumption that public schools as currently constituted fail to transform inputs into outputs. Before profound changes in schools are made because of a presumed — and apparently inaccurate — conclusion that resources are unrelated to achievement, compelling evidence of the efficacy of the proposed changes should be required.

I. Reanalysis of Hanushek's literature review

To enable this reanalysis, Eric Hanushek provided the classification of estimates and studies underlying his 1997 literature summary.[6] As he writes (1997, 142),

> This summary concentrates on a set of published results available through 1994, updating and extending previous summaries (Hanushek 1981, 1986, 1989). The basic studies meet minimal criteria for analytical design and reporting of results. Specifically, the studies must be published in a book or journal (to ensure a minimal quality standard), must include some measures of family background in addition to at least one measure of resources devoted to schools, and must provide information about statistical reliability of the estimates of how resources affect student performance.

Hanushek describes his rule for selecting estimates from the various studies in the literature as follows:

> The summary relies on all of the separate estimates of the effects of resources on student performance. For tabulation purposes, a "study" is a separate estimate of an educational production found in the literature. Individual published analyses typically contain more than one set of estimates, distinguished by different measures of student performance, by different grade levels, and frequently by entirely different sampling designs.

Most of the studies underlying Hanushek's literature summary were published in economics journals.

Table 1-1 summarizes the distribution of the estimates and studies underlying Hanushek's literature tabulation. The first column reports the number of estimates used from each study, dividing studies into those where only one estimate was used (first row), two or three were used (second row), four to seven were used (third row), or eight or more were used (fourth row). Seventeen studies contributed only one estimate each,[7] while nine studies contributed eight or more estimates each. These latter nine studies made up only 15% of

TABLE 1-1
Distribution of class size studies and estimates taken in Hanushek (1997)

Number of estimates used (1)	Number of studies (2)	Number of estimates contributed (3)	Percent of studies (4)	Percent of estimates (5)
1	17	17	28.8%	6.1%
2-3	13	28	22.0	10.1
4-7	20	109	33.9	39.4
8-24	9	123	15.3	44.4
Total	59	277	100.0%	100.0%

Note: Column (1) categorizes the studies according to the number of estimates that were taken from the study. Column (2) reports the number of studies that fall into each category. Column (3) reports the total number of estimates contributed from the studies. Column (4) reports the number of studies in the category as a percent of the total number of studies. Column (5) reports the number of studies in the category as a percent of the total number of estimates used from all the studies.

the total set of studies, yet they contributed 44% of all estimates used. By contrast, the 17 studies from which only one estimate was taken represented 29% of studies in the literature and only 6% of the estimates. The studies providing only a single estimate tend to be based on larger samples, providing more accurate estimates. There is also good reason to think that studies with negative or insignificant results will tend to contain more estimates. Therefore, Hanushek's estimate selection scheme has the effect of over-weighting imprecise and unsystematic estimates relative to more reliable ones.

A consideration of Hanushek's classification of some of the individual studies in the literature helps to clarify his procedures. Two studies by Link and Mulligan (1986 and 1991) each contributed 24 estimates, or 17% of all estimates. Both papers estimated separate models for math and reading scores by grade level (third, fourth, fifth, or sixth) and by race (black, white, or Hispanic), yielding 2 x 4 x 3 = 24 estimates apiece. One of these papers, Link and Mulligan (1986), addressed the merits of a longer school day using an 8% subsample of the dataset used in the other paper (1991). Class size was not the focus of this paper, and it was included in the regression specifications only in an interaction with peer ability levels. In a passing statement, Link and Mulligan (1986, 376) note that, when they included class size separately in their 12 equations for the math score, it was individually statistically insignificant.[8] Link and Mulligan (1991), which concentrated on estimating the impact of peer group effects on student achievement, did not explicitly control for family background in any of its estimates, although separate equations were estimated for black, white, and Hispanic students.

By contrast, Card and Krueger (1992) focused on the effect of school resources on the payoff from attending school longer, and presented scores of estimates for 1970 and 1980 Census samples of white males sometimes exceeding one million observations (see, e.g., their Table 6). Nonetheless, Hanushek informed me he extracted only one estimate from this study because only one specification controlled explicitly for family background information, although all the estimates conditioned on race in the same fashion as Link and Mulligan's (1991) 24 estimates.[9]

No estimates were selected from Finn and Achilles's (1990) analysis of the STAR experiment.

TABLE 1-2
Reanalysis of Hanushek's (1997) literature summary of class size studies

Result	Hanushek weights (1)	Equally weighted studies (2)	Weighted by number of citations (3)	Selection-adjusted weighted studies (4)
Positive and stat. sig.	14.8%	25.5%	30.6%	33.5%
Positive and stat. insig.	26.7	27.1	21.1	27.3
Negative and stat. sig.	13.4	10.3	7.1	8.0
Negative and stat. insig.	25.3	23.1	26.1	21.5
Unknown sign and stat. insig.	19.9	14.0	15.1	9.6
Ratio positive to negative	1.07	1.57	1.56	2.06
P-value*	0.500	0.059	0.096	0.009

Note: See text for full explanation. Column (1) is from Hanushek (1997, Table 3), and implicitly weights studies by the number of estimates that were taken from each study. Columns (2), (3), and (4) are author's tabulations based on data from Hanushek (1997). Column (2) weights each estimate by one over the number of estimates taken from that study, thus weighting each study equally. Column (3) calculates a weighted average of the data in column (2), using the number of times each study was cited as weights. Column (4) uses the regressions in Table 1-3 to adjust for sample selection (see text). A positive result means that a smaller class size is associated with improved student performance. The table is based on 59 studies.

* P-value corresponds to the proportion of times the observed ratio, or a higher ratio, of positive to negative results would be obtained in 59 independent random draws in which positive and negative results were equally likely.

Hanushek informed me that this decision was made because Finn and Achilles did not control for family background (other than race and school location). However, the STAR experiment used random assignment of students to classes, and econometric reasoning suggests that controls for family background should therefore be unnecessary (because family background variables and class size are expected to be uncorrelated).

Column 1 of **Table 1-2** summarizes Hanushek's tabulation of the estimates he selected from the literature. His approach equally weights all 277 estimates drawn from the underlying 59 studies. Following Hanushek, estimates that indicate that smaller classes are associated with better student performance are classified as positive results.[10] The bottom of the table reports the ratio of the number of positive to negative results. Below this is the p-value that corresponds to the probability of observing so high a ratio if, in fact, there were no relationship between class size and student performance and each study's results were merely a random draw with positive and negative results equally likely.[11] That is, how different are the results from a series of coin flips in which positive (heads) or negative (tails) results are equally likely in each study? A p-value of less than 0.05 indicates that the observed ratio of positive to negative results would occur by chance less than one time in 20, and is typically taken as evidence of a statistically significant relationship. Column 1, with a p-value of 0.500, indeed shows no systematic relationship between smaller classes and better student performance; estimates are virtually equally likely to be negative as positive. Only one quarter of the estimates are statistically significant, and these are also about equally likely to be negative as positive.

As mentioned, Hanushek's procedure places more weight on studies from which he extracted more estimates. There are a number of reasons to question the statistical properties of such an approach. First, studies that contain many estimates are likely to have broken their data into several subsamples, and as a result estimates based on subsamples are given extra weight. These estimates by definition have fewer

observations — and higher sampling variances — than estimates based on the full samples, and an optimal weighting scheme should therefore give them *lower* weights.[12] Second, there is reason to suspect a systematic relationship between a study's findings and the number of estimates it contains. Most people expect there to be a positive relationship between small classes and test performance. Authors who find weak or negative results (e.g., because of sampling variability or specification errors) may be required by referees to provide additional estimates to probe their findings (or they may do so voluntarily), whereas authors who use a sample or specification that generates an expected positive effect may devote less effort to reporting additional estimates for subsamples. If this is the case, and if findings are not independent across estimates (as would be the case if a mis-specified model is estimated on different subsamples), then Hanushek's weighting scheme will place too much weight on insignificant and negative results.

Figure 1A provides evidence that Hanushek's procedure assigns excessive weight to studies with unsystematic or negative results. The figure shows the fraction of estimates that are positive, negative, or of unknown sign, by the number of estimates Hanushek took from each study. For the vast majority of studies from which Hanushek took only a small number of estimates, there is a clear and consistent association between smaller class sizes and student achievement. In the 17 studies from which Hanushek took only one estimate, for example, more than 70% of the estimates indicate that students tend to perform better in smaller classes while only 23% indicate a negative effect. By contrast, in the nine studies from which Hanushek took eight or more estimates each — for a total of 123 estimates — the opposite pattern holds: small classes are more likely to be associated with lower performance.

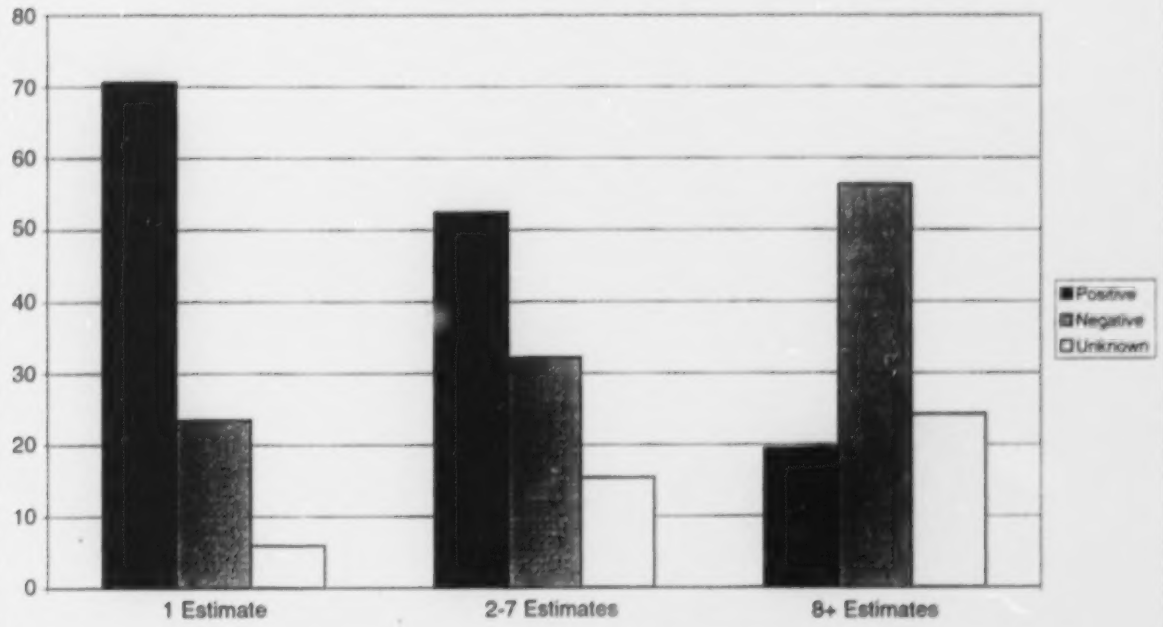
Table 1-3 more formally explores the relationship between the number of estimates that Hanushek extracted from each study and their results. Specifically, column 1 reports a bivariate regression in which the dependent variable is the percent of estimates in a study that are positive and statistically significant (based on Hanushek's classification) and the explanatory variable is the number of estimates that Hanushek took from the study. The unit of observation in the table is a study, and the regression is estimated for Hanushek's set of 59 studies. Columns 2-5 report analogous regressions where the dependent variable is the percent of estimates that are positive and insignificant, negative and significant, negative and insignificant, or of unknown sign, respectively. These results show that Hanushek's summary uses fewer estimates from studies that tended to find positive and significant results ($r = -0.28$), and this relationship is stronger than would be expected by chance alone. Moreover, the opposite pattern holds for studies with negative and significant findings: relatively more estimates from studies with perverse class size effects are included in the sample, although this relationship is not significant. Table 1-3, then, seems to provide strong evidence that Hanushek's selection criteria have the effect of biasing his representation of the literature toward finding zero or negative effect of class size on performance.

The rule that Hanushek used for selecting estimates would be expected to induce a positive association between the prevalence of insignificant results and the number of estimates taken from a study, since studies with more estimates probably used smaller subsamples (which are more likely to generate insignificant estimates). *But Table 1-3 also shows something different: that Hanushek took more estimates from studies that had negative, statistically significant results. Sampling bias resulting from smaller subsamples cannot explain this,* although one explanation may come from the refereeing process discussed above. In any case, given this aspect of Hanushek's estimate selection process, we should expect his results to be biased toward a negative or unsystematic effect of class size reduction; it is not surprising that he found little evidence for a positive effect.

The remaining columns of Table 1-2 attempt to remove the bias from Hanushek's procedure by weight-

FIGURE 1A

Average percent of estimates positive, negative, or unknown sign, by number of estimates taken from study



Notes: Based on data from Hanushek (1997). Arithmetic averages of percent positive, negative, and unknown sign are taken over the studies in each category

TABLE 1-3

Regressions of percent of estimates positive or negative, and significant or insignificant, on the number of estimates used from each study

	Dependent variable:				
	Percent positive & significant (1)	Percent positive & insignificant (2)	Percent negative & significant (3)	Percent negative & insignificant (4)	Percent unknown sign & insignificant (5)
Intercept	35.7	27.4	7.4	21.0	8.5
	(6.4)	(6.0)	(4.5)	(5.9)	(5.6)
Number of estimates used	-2.16	-0.07	0.62	0.44	1.18
	(0.96)	(0.89)	(0.66)	(0.88)	(0.83)
R-square	0.08	0.00	0.01	0.00	0.03

Notes: Standard errors are shown in parentheses. Sample size is 59 studies. Dependent variable is the percent of estimates used by Hanushek in each result category. Unit of observation is a study.

ing the different studies more appropriately. As a partial correction for the oversampling from studies with negative and insignificant estimates, in column 2 of Table 1-2 the underlying studies — as opposed to the individual estimates extracted from the studies — are given equal weight. This is accomplished by assigning to each study the percent of estimates that are positive and significant, positive and insignificant, and so on, and then taking the arithmetic average of these percentages over the 59 studies.[13] This simple and plausible change in the weighting scheme substantially alters the inference one draws from the literature. In particular, studies with positive effects of class size are 57% more prevalent than studies with negative effects.

In column 3 of Table 1-2 an alternative approach is used. Instead of weighting the studies equally, studies are weighted based on a measure of their quality, as indicated by the frequency with which they are cited. Studies are assigned a weight equal to the cumulative number of citations to the study as of August 1999, based on a "cited reference search" of the *Social Science Citation Index*. Column 3 presents the weighted mean of the percentages. Although there are obvious problems with using citations as an index of study quality (e.g., articles published earlier have more opportunity to be cited; norms and professional practices influence the number of citations, etc.), citation counts are a widely used indicator of quality, and should be a more reliable measure of study quality than the number of estimates Hanushek extracted. The results are similar to those in column 2: studies with statistically significant, positive findings outweigh those with statistically significant, negative findings by over 2 to 1.

Another alternative, and in some respects superior, approach to adjust for estimate selection bias is to use the regressions in Table 1-3 to generate predicted percentages for all studies under the hypothetical situation in which one estimate was provided by each study. This is akin to creating a simulated dataset that looks like Hanushek's data might have looked if he took only one estimate from each study. This approach would be preferable to the equally-weighted-studies approach in column 2 if the primary estimate in each study tends to be systematically different from the secondary estimates. Such a pattern could arise, for example, if the first estimate that each study presents is for its full sample, and subsequent estimates carve the sample into smaller subsamples that naturally yield noisier estimates. A linear approximation to what the average study would find if one estimate were extracted from all studies is derived by adding together the intercept and slope in each of the regression models in Table 1-3. These results predict what the outcome would have been if each study had reported only one estimate.[14] Column 4 of Table 1-2 reports the distribution of results using this simulated dataset. This approach for adjusting for the selection of estimates from the studies indicates even stronger and more consistent positive effects of class size. After adjusting for selection, studies with positive results are twice as likely as studies with negative results; if in fact there were no positive relationship between performance and small classes, the probability of observing this many studies with positive results by chance would be less than one in a hundred. Among studies with statistically significant results, positive results outnumber negative results by 4 to 1.

In sum, all three of these alternatives to Hanushek's weighting scheme produce results that point in the opposite direction of his findings: all three find that smaller class sizes are positively related to performance, and that the pattern of results observed in the 59 studies is unlikely to have arisen by chance. It should be emphasized that the results reported in Table 1-2 are all based on Hanushek's coding of the underlying studies. Although Hanushek (1997) tried to "collect information from all studies meeting" his selection criteria, he notes that "[s]ome judgment is required in selecting from among the alternative specifications." *The selection and classification of estimates in many of the studies is open to question, and could in part account for the curious relationship between the number of estimates taken from a study and the study's findings.* The following examples illustrate some additional types of problems encountered in the way studies were coded, and the limitations of some of the underlying estimates:

- As mentioned previously, the Link and Mulligan (1986) study was classified as having 24 statistically insignificant estimates of unknown sign, although the authors mention that class size was insignificant in only 12 of the equations they estimated, use a subsample of a larger dataset also used in another paper, and do not report tests for the joint significance of class size and peer group achievement (which typically indicate that smaller classes have beneficial effects). The median sample size in this paper was 237, compared with 3,300 in Link and Mulligan (1991), yet all estimates received equal weight.

- Kiesling (1967) was classified as having three estimates of the effect of class size, but there is no mention of a class size variable in Kiesling's paper.

- Burkhead's (1967) study yielded 14 estimates, all of which were statistically insignificant (three quarters were negative). Four of these estimates are from a sample of just 22 high-school-level observations in Atlanta.[15] Moreover, the outcome variable in some of the models, post-high-school-education plans, was obtained by "a show of hands survey in the high schools." Despite these limitations, with 14 estimates this study receives over three times as much weight as the median study in Hanushek's summary.

- At least a dozen of the studies that Hanushek included in his sample estimated regression models that included expenditures per pupil and teachers per pupil as separate regressors in the same equation (e.g., Maynard and Crawford 1976). The interpretation of the teachers-per-pupil variable in these equations is particularly problematic because one would expect the two variables (expenditures per pupil and teachers per pupil) to vary together. One can identify the separate effect of teachers per pupil only if they do not vary together, which is most likely to happen when there are differences between schools in teacher salaries. That is, if School A has a lower pupil-teacher ratio than School B, but the schools have equal expenditures per pupil, the most likely way School A achieved a lower pupil-teacher ratio is by paying its teachers less — a difference that obviously could influence student achievement.[16] Using this source of variability in class size obviously changes the interpretation of the class size result, and renders the finding irrelevant for most policy considerations.

Expenditures per student

Hanushek (1997) also examines the effect of expenditures per student, although he argues that "studies involving per-pupil expenditure tend to be the lowest quality studies." **Table 1-4** is analogous to Table 1-2 for the expenditure-per-pupil studies. The first column uses Hanushek's method, which weights studies by the number of estimates he extracted from them. The second column equally weights each study. The third column weights the studies by the number of times the article has been cited, and the fourth column uses the regression-adjustment method described above. In all cases, the relative frequency of studies that find positive effects of expenditures per student is greater than would be expected by chance. A total of 163 estimates were extracted from 41 studies.

The following regression coefficients describe the relationship between the percent of estimates that are positive, negative, or of unknown sign, and the number of estimates represented by the study, for the 41 studies in Hanushek's summary. (Standard errors for the coefficients are in parentheses, and an asterisk indicates a statistically significant coefficient at the 0.10 level.)

TABLE 1-4
Reanalysis of Hanushek's (1997) literature summary; studies of expenditures per pupil

Result	Hanushek weights (1)	Equally weighted studies (2)	Weighted by number of citations (3)	Selection-adjusted weighted studies (4)
Positive and stat. sig.	27.0%	38.0%	33.5%	50.5%
Positive and stat. insig.	34.3	32.2	30.5	29.7
Negative and stat. sig.	6.7	6.4	2.7	6.0
Negative and stat. insig.	19.0	12.7	14.8	5.5
Unknown sign and stat. insig.	12.9	10.7	18.4	8.3
Ratio positive to negative	2.39	3.68	3.66	6.97
P-value*	0.0138	0.0002	0.0002	0.0000

Notes: Column (1) is from Hanushek (1997, Table 3), and implicitly weights studies by the number estimates that were taken from each study. Columns (2), (3), and (4) are author's tabulations based on data from Hanushek (1997). Column (2) assigns each study the fraction of estimates corresponding to the result based on Hanushek's coding, and calculates the arithmetic average. Column (3) calculates a weighted average of the data in column (2), using the number of times each study was cited as weights. Column (4) uses regressions corresponding to Table 1-3 to adjust for sample selection (see text). A positive result means that a smaller class size is associated with improved student performance. The table is based on 41 studies.

* P-value corresponds to the proportion of times the observed ratio, or a higher ratio, of positive to negative results would be obtained in 41 independent Bernouli trials in which positive and negative results were equally likely.

Percent positive = 83.6* - 3.4* (number of estimates used) $R^2 = .09$
(8.9) (1.7)

Percent negative = 8.9 + 2.6* (number of estimates used) $R^2 = .08$
(6.9) (1.3)

Percent unknown = 7.5 + 0.8 (number of estimates used) $R^2 = .01$
(7.6) (1.5)

As with the class size studies, Hanushek extracted more estimates from studies that tended to find insignificant or negative effects of expenditures per student and fewer from studies that found positive effects. The dependence between the number of estimates and a study's results accounts for why Hanushek's technique of weighting more heavily the studies from which he took more estimates produces the least favorable results for expenditures per student. All of the various weighting schemes in Table 1-4 indicate that greater expenditures are associated with greater student achievement.

Summing up

In response to work by Hedges, Laine, and Greenwald (1994), Hanushek (1996b, 69) argued that, "[u]nless one weights it in specific and peculiar ways, the evidence from the combined studies of resource usage provides the answer" that resources are unrelated to academic achievement, on average. Since Hanushek's results are produced by implicitly weighting the studies by the *number* of "separate" estimates they present (or more precisely, the number of estimates he extracted from the studies), it seems likely that the opposite conclusion is more accurate: unless one weights the studies of school resources in peculiar ways, the *average study* tends to find that more resources are associated with greater student achievement.

This conclusion does not, of course, mean that reducing class size is necessarily worth the additional investment, or that class size reductions benefit all students equally. These questions require knowledge of the strength of the relationships between class size and economic and social benefits, knowledge of how these relationships vary across groups of students, and information on the cost of class size reduction. These issues are taken up in the next section. But the results of this reanalysis of Hanushek's literature summary should give pause to those who argue that radical changes in public school incentives are required because schooling inputs are unrelated to schooling outputs. When the study is the unit of observation, Hanushek's coding of the literature suggests that class size is a determinant of student achievement, at least on average.

II. Economic criterion

Hanushek (1997, 144) argues that, "[g]iven the small confidence in just getting noticeable improvements [from school resources], it seems somewhat unimportant to investigate the size of any estimated effects." This argument is unpersuasive for at least two reasons. First, as argued above, Hanushek's classification of studies in the literature indeed provides evidence of a systematic relationship between school inputs and student performance for the typical school district. Second, if the estimates in the literature are imprecise (i.e., have large sampling variances), statistically insignificant estimates are not incompatible with large economic and social returns from reducing class size. The power of the estimates is critical: if a given study cannot statistically distinguish between a large positive effect of reducing class size and zero effect, it tells us little about the value of class size reductions. Statistical significance tells us only whether a zero effect can be rejected with confidence. But zero is not a very meaningful null hypothesis in this case: we would also be reluctant to spend large amounts of money to reduce class sizes if the effect on outcomes was positive but *small*. What would be a more meaningful null hypothesis? One way to approach this question is to estimate a break-even point — the minimum benefit to reducing class size that would justify its cost — and use this as a basis for comparison. This section provides calculations suggesting a reasonable null hypothesis for the effect of class size based on standard economic considerations, and compares this to the results of the STAR experiment.

Lazear's theory of class size

A recent paper by Edward Lazear (1999) lays out an insightful economic theory of class size. In essence, Lazear argues that students who attend a smaller class learn more because they experience fewer student disruptions during class time, on average. Such a result follows naturally if the probability of a child disrupting a class is independent across children. Lazear then quite plausibly assumes that disruptions require teachers to suspend teaching, creating a "negative externality" that reduces the amount of learning for everyone in the class. There may be other benefits to smaller classes as well. For example, it is possible that students who spend time in small classes learn to behave better with closer supervision, leading to a reduced propensity to disrupt subsequent classes. Lazear's model probably captures an important feature of class size, and yields a specific functional form for the education production function.

Another implication of Lazear's model is that the "optimal" class size is larger for groups of students who are well behaved, because these students are less likely to disrupt the class and therefore benefit less from a class size reduction than more disruptive students. Schools therefore have an incentive to assign weaker, more disruptive students to smaller classes. Compensatory education programs that provide more resources to lower-achieving schools could also be viewed as targeting resources to weaker students. If schools voluntarily assign weaker students to smaller classes (as predicted by Lazear) or if compensatory

funding schemes cause weaker students to have smaller classes, a spurious negative association between smaller classes and student achievement would be created. This phenomenon could explain why studies that avoid this problem by focusing on changes in class size that are not chosen by school administrators but are imposed from outside for reasons unrelated to individual students — such as in Angrist and Lavy's (1999) clever analysis of Israel's Maimonides law, as well as the STAR experiment — tend to find that smaller classes have a beneficial effect on student achievement. For educational policy, the relevant parameter is the potential gain in achievement from exogenous reductions in class size from current levels, not the relationship estimated from observed variations in class sizes voluntarily chosen by schools.

One final aspect of Lazear's model is worth emphasizing. If schools behave optimally, then they will reduce class size to the point that the benefit of further reductions in class size just equals the cost.[17] This implication provides a plausible economic null hypothesis. If we are starting from the optimal level, the costs and benefits of changes in class size should be roughly equivalent. As Lazear (1999) writes, "The point is that even if class size effects are potentially important, in equilibrium, marginal changes in class size may have small effects on observed educational output. If large gains were available from lowering class size, then those changes would have been made." Unless large opportunities for social gain are left unexploited by local school districts, we would expect the benefits of further reductions in class size to equal their costs.

Benefits and costs of educational resources

Improved school resources can have many benefits for students. This section focuses on one particular potential benefit: the effect on the students' future labor market earnings. Improved school resources might help students learn more and, separately, raise their educational aspirations. These can both pay off in the labor market, leading to better job placements and higher earnings within each job. This section attempts to quantify the size of this benefit by combining the effect of school resources on standardized test scores with the relationship between test scores and labor market earnings.

Several studies have examined the relationship between students' test scores while in school and their subsequent earnings. Three recent studies illustrate the magnitude of this relationship:

- Murnane, Willet, and Levy (1995), using data from the High School and Beyond survey, estimate that male high school seniors who scored one standard deviation (SD) higher on the basic math achievement test in 1980 earned 7.7% higher earnings six years later, and females earned 10.9% more. This study, however, also controls for students' eventual educational attainment, so any effect of test scores on educational attainment — which, of course, affects wages — is not attributed to the influence of test scores.

- Currie and Thomas (1999) use the British National Child Development Study to examine the relationship between math and reading test scores at age 7 and earnings at age 33. They find that students who score in the upper quartile of the reading exam earn 20% more than students who score in the lower quartile of the exam, while students in the top quartile of the math exam earn another 19% more.[18] Assuming that scores are normally distributed, the average student in the top quartile scores about 2.5 standard deviations higher than the average student in the bottom quartile, so these results imply that a one SD increase in reading test performance is associated with 8.0% higher earnings, while a one standard deviation increase in the math test is associated with 7.6% higher earnings.

- Neal and Johnson (1996) use the National Longitudinal Survey of Youth to estimate the effect of stu-

dents' scores on the Armed Forces Qualification Test (AFQT), taken at age 15-18, on their earnings at age 26-29. Adjusting for the students' age when the test was taken, they find that a one SD increase in scores is associated with about 20% higher earnings for both men and women.

There are probably three important reasons why Neal and Johnson find a larger effect of test scores on wages than do Currie and Thomas. First, Currie and Thomas use a test administered at age 7, while Neal and Johnson use a test that was administered when their sample was in its late teens. Currie and Thomas find some mean regression in test scores — students who score very high at young ages tend to have smaller score increases as they age than do students who score very low on the earlier test — which suggests that a later test might be a stronger predictor of earnings. Second, Neal and Johnson use only a single test score while Currie and Thomas use both reading and math scores, which are correlated. Finally, differences between British and American labor markets might account for part of the difference. Based on these three studies, a plausible assumption is that a one SD increase in either math or reading scores is associated with about 8% higher earnings.

From an investment perspective, the timing of costs and benefits is critical. The costs of hiring additional teachers and obtaining additional classrooms are borne up front, while the benefits are not realized until years later, after students join the labor market. Delayed benefits need to be discounted to make them comparable to upfront costs. To illustrate the benefits and costs, consider extending the STAR class size reduction experiment to the average U.S. student entering kindergarten in 1998. In the STAR experiment, classes were reduced from about 22 to about 15 students, so assume that funds are allocated to create 47% (7/15) more classes.

Probably a reasonable approximation is that the cost of creating and staffing more classrooms is proportional to the annual per pupil cost.[19] We assume for this cost-benefit calculation that the additional cost per pupil each year a pupil is in a small class equals $3,501, or 47% of $7,502, which was the nationwide total expenditures per student in 1997-98.[20] Although the STAR experiment lasted four years, the average student who was assigned to a small class spent 2.3 years in a small class.[21] We also assume the additional costs are $3,501 in years 1 and 2, 30% of $3,501 in year 3, and zero in year 4. Denote the cost of reducing class size in year t as C_t. The present value (PV) of the costs discounted to the initial year (1998) using a real discount rate of r is:

$$PV\ of\ Costs = \sum_{t=1}^{4} C_t/(1+r)^t.$$

Column 2 of **Table 1-5** provides the present value of the costs for various values of the discount rate.

The economic benefits of reduced class size are harder to quantify, and occur further in the future. Suppose initially that the earnings of the current labor force represent the exact age-earnings profile that the average student who entered kindergarten in 1998 will experience when he or she completes school and enters the labor market. **Figure 1B** illustrates this age-earnings profile for workers in 1998.[22] The figure displays average annual earnings for workers at each age between 18 and 65. As is commonly found, earnings rise with age until workers reach the late 40s, peak in the early 50s, and then decline. Average earnings are quite low until workers reach their mid-20s. Let E_t represent the average real earnings each year after age 18.

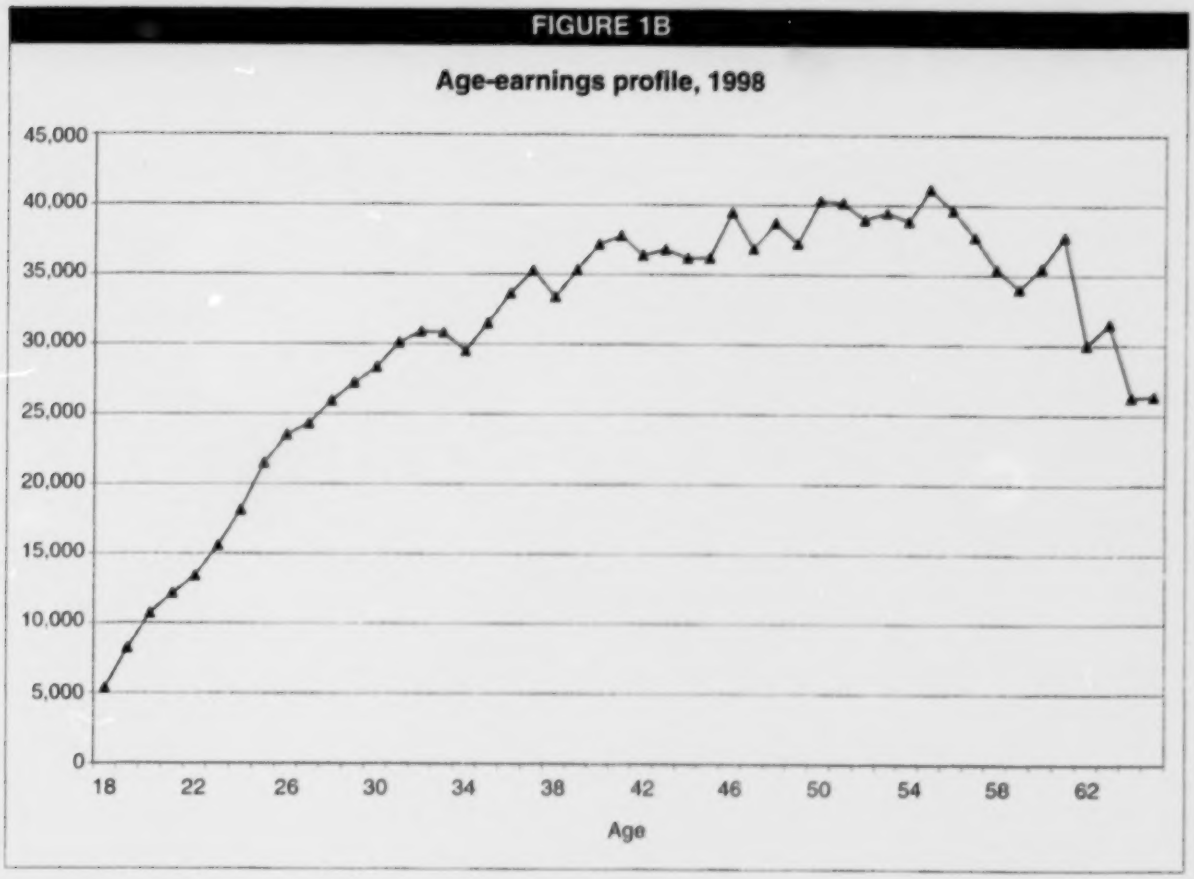
Assume that ß represents the increase in earnings associated with a one standard deviation increase in either math or reading test scores. The preceding discussion suggests that 8% is a reasonable estimate for the value of ß. Now let δ_M and δ_R represent the increase in math and reading test scores (in SD units) due to attending smaller classes in grades K-3. The STAR experiment suggests that $\delta_M = \delta_R = 0.20$ SD is a reason-

TABLE 1-5
Discounted present value of benefits and costs of reducing class size from 22 to 15 in grades K-3 (1998 dollars)

Discount rate (1)	Cost (2)	Increase in income assuming annual productivity growth rate of: None (3)	1% (4)	2% (5)
0.02	$7,787	$21,725	$31,478	$46,294
0.03	7,660	15,174	21,667	31,403
0.04	7,537	10,784	15,180	21,686
0.05	7,417	7,791	10,819	15,238
0.06	7,300	5,718	7,836	10,889

Note: Figures assume that a one standard deviation increase in math test scores or reading test scores in grades K-3 is associated with an 8% increase in earnings, and that attending a small class in grades K-3 raises math and reading test scores by 0.20 SD. Real wages are assumed to grow at the same rate as productivity. Costs are based on the assumption that students are in a smaller class for 2.3 years, as was the average in the STAR experiment.

FIGURE 1B



able figure to use (see, e.g., Finn and Achilles 1990 or Krueger 1999b). Then the average real earnings of students from smaller classes is E_t x $(1+\beta(\delta_M + \delta_R))$. This exceeds average real earnings of students from regular-size classes by E_t x $\beta(\delta_M + \delta_R)$. The addition to annual earnings must be discounted back to the initial year to account for the fact that a dollar received in the future is less valuable than a dollar received today. Assuming students begin work at age 18 and retire at age 65, the appropriate formula for discounting the higher earnings stream due to smaller classes back to the beginning of kindergarten is:

$$PV\ of\ Benefits = \sum_{t=14}^{61} E_t x\beta(\delta_M + \delta_R) / (1 + r)^t.$$

Using these assumptions, column 3 of Table 1-5 reports the present value of the additional earnings due to reducing class size by seven students for various values of the discount rate.

One important issue, however, is that real average earnings are likely to grow substantially between 1998 and the year when the average kindergartner of 1998 retires. That is, when the kindergartners of 1998 enter the labor market, their average earnings (after adjusting for inflation) will be greater than that depicted in Figure 1B. Real wages typically grow in step with labor productivity (i.e., output per hour). Over the 20th century, real earnings and productivity have typically grown by 1% or 2% per year. The estimates of ß discussed above are all based on earnings long after students started school, which reflect the effect of higher productivity growth on earnings. Consequently, columns 4 and 5 present discounted benefits assuming either 1% or 2% annual productivity and real wage growth after 1998.[23] The latest Social Security Trustees' intermediate projection is for real wages to grow by slightly less than 1% per year over the next 75 years, so column 4 arguably probably provides a reasonable forecast of future earnings.

The next question is, which discount rate should one use to discount costs and benefits from age 5 until 65? The current yield on essentially risk-free long-term inflation-indexed government bonds is just under 4%. If we assume an interest rate of 4% (row 3), then the benefits of reducing class size from 22 to 15 in the early grades would be 43% greater than the costs absent real wage growth, and 100% greater than the costs if real wages grow by 1% per year. However, because the payoff to reduced class sizes is uncertain, society might desire to reflect some risk in the interest rate used to discount future benefits. A higher discount rate would then be desired. With a discount rate of 6% and 1% annual productivity growth, the costs of reducing class size from 22 to 17 students are predicted to almost equal the benefits, in line with Lazear's prediction.

An informed reader might question whether a 0.20 standard deviation gain from smaller classes is appropriate for the calculations in Table 1-5. In particular, work by Krueger and Whitmore (1999) and Nye, Zaharias, Fulton, et al. (1994) suggests that the improved test performance of small-class students in Project STAR may have fallen to about 0.10 standard deviations by the end of high school.[24] Although I suspect that some of the initial gain from small classes in the STAR experiment faded after students returned to regular-size classes, the calculations reported in Table 1-5 are probably still reasonable. The reason for this supposition is that Currie and Thomas's estimate of ß is based on test scores at age 7. They find some regression to the mean in test scores as students age — that is, students with high scores at age 7 tend to drift lower, while students with low initial scores tend to see larger increases. This regression to the mean is consistent with the Krueger and Whitmore and Nye et al. results mentioned above. If the 0.10 SD gain in test scores at older ages is to be used in the calculations, then a higher value of ß would be appropriate, as test scores of high school seniors are more strongly correlated with eventual earnings than are those students' scores at age 7.

The 'critical effect size'

Another, perhaps more relevant, way to consider the benefit-cost calculus is to ask, "What is the minimum increase in test scores from a reduction in class size of seven students in grades K-3 that is required to justify the added cost?" That is, at what size of the increase in test scores do the benefits of class size reduction exactly equal the costs? This "critical effect size" provides a logical null hypothesis for policy makers and researchers to use in evaluating the economic significance of the class size literature. The critical effect size was calculated by solving for δ^* in the following equation:

$$\sum_{t=1}^{4} C_t/(1+r)^t = \sum_{t=14}^{61} E_t x(.08)(2\delta^*)/(1+r)^t,$$

where math and reading scores are assumed to increase by the same amount due to smaller classes, and β has been fixed at 0.08.

Estimates of the "critical effect size" for various values of the discount rate and productivity growth are reported in **Table 1-6**. A noteworthy finding is that the critical effect size is fairly small. If we use a 4% discount rate and expect 1% annual productivity growth, the minimum increase in elementary school math and reading scores required for the benefits to equal the costs of a class size reduction from 22 to 15 students is 0.10 standard deviations. The critical effect size for a class size reduction of one student, from 22 to 21, would be 0.010 standard deviations.[25] I suspect that most of the estimates in the literature would have difficulty rejecting a critical effect size of this magnitude. Unfortunately, most studies do not provide sufficient information to test their results against this hypothesis.

Also notice that the effect sizes found in the STAR experiment and much of the literature are greater for minority and economically disadvantaged students than for other students. Although the critical effect size differs across groups with different average earnings, economic considerations suggest that resources would be optimally allocated if they were targeted toward those who benefit the most from smaller classes.

Caveats

Many assumptions underlying the cost-benefit calculations in Tables 1-5 and 1-6 could turn out to be wrong. The assumptions that are probably most critical are:

- The effect of test score gains on earnings in the future may turn out to be different than the value of ß that was assumed. Indeed, because β was estimated from cross-section relations, it could reflect the effect of omitted characteristics, which would imply that it does not reflect the potential gain from increasing a particular student's scores.[26] In addition, general equilibrium effects could affect the value of ß if class size is reduced on a wide scale — a substantial increase in the cognitive ability of the labor force would be expected to reduce the return to cognitive ability. It is also likely that school resources influence earnings by means that are not reflected in test scores. For example, class size may influence non-cognitive abilities, which are not reflected in test scores but nevertheless influence earnings, especially for blue-collar workers (see Cawley et al. 1996).

- Class size probably influences other outcomes with economic consequences, such as crime and welfare dependence, and there may be externalities from human capital, so the economic benefits could be understated. There are also non-economic benefits of improved education. None of these are captured by the focus here on individual earnings.

TABLE 1-6
Required standard deviation increase in elementary school math and reading test scores for a class size reduction of seven students to break even

Discount rate (1)	Critical effect size assuming annual productivity growth rate: None (2)	1% (3)	2% (4)
0.02	0.072	0.049	0.034
0.03	0.101	0.071	0.049
0.04	0.140	0.099	0.070
0.05	0.190	0.137	0.097
0.06	0.255	0.186	0.134

Note: Figures assume that a one standard deviation increase in math test scores or reading test scores in grades K-3 is associated with an 8% increase in earnings. Real wages are assumed to grow at the same rate as productivity.

- It is unclear how much real earnings will grow in the future, although the 0-2% annual growth figures probably provide a reasonable range.
- The cost of reducing class size in the early grades may be different than assumed here. For example, expenditures per student are typically lower in grammar school, yet the analysis here uses expenditures per student in all grades as the basis for calculations. More importantly, the STAR experiment reduced only the number of classroom teachers, whereas the calculations here assume an across-the-board reduction in the number of teachers (including, e.g., physical education teachers, music teachers, and art teachers). Furthermore, the existence of fixed costs (e.g., administration, transportation) would also cause my assumption that costs are proportional to the number of teachers per pupil to overstate costs. These considerations suggest that the costs of class size reduction assumed here may have been substantially overstated.
- If class size is to be reduced on a wide scale, a great many new teachers will be needed to teach the new classes. In the short run, this could cause the quality of teachers to decline. On the other hand, more qualified individuals may be attracted to the teaching profession if classes are smaller.
- The calculations on workers' earnings in Tables 1-5 and 1-6 neglect fringe benefits, which are about 20% of total compensation. If fringe benefits increase in proportion to earnings, the reported benefits are understated by about 20%. The calculations also assume that everyone works for pay, at least part year, which tends to overstate the economic benefit, probably by 20% or so.

The related literature that directly examines the effect of expenditures per student on students' subsequent income provides some independent support for the calculations underlying Tables 1-5 and 1-6. Card and Krueger (1996) review 11 such studies. Although these studies are less well controlled than the STAR experiment, the median estimate is that a 10% increase in expenditures per student is associated with 1.25% higher earnings, and the inter-quartile range — the difference between the 75th percentile study and the 25th percentile study — is from 0.85% to 1.95%.[27] It turns out that this is quite close to the estimate derived above

using the STAR experiment: if we assume linearity and β=0.08, then the STAR experiment implies that a 10% reduction in class size leads to a 1.0% increase in earnings.[28] Thus, despite employing quite different estimation procedures, the literature that directly estimates the effect of class size on earnings yields results that are in the same ballpark as the corresponding figure derived from the STAR experiment.

III. Conclusion

The method Hanushek uses to summarize the literature is often described as a "vote counting" exercise. The results are shown to depend critically on whether the approach allows *one study, one vote.* When studies are accorded equal weight, the literature exhibits systematic evidence of a relationship between class size and achievement. As implemented by Hanushek, however, studies from which multiple estimates were extracted are given multiple votes. No statistical theory is presented to support this weighting scheme, and it can be misleading. There are good reasons to think that this scheme leads to over-weighting studies with less systematic and less significant estimates. For example, other things equal, studies that report a larger number of estimates for finer subsamples will tend to yield less significant estimates, but will be given extra weight by Hanushek's weighting scheme. Studies are a more natural unit of observation, as it is studies, not estimates, that are accepted for publication. The importance of a study as the unit of observation is acknowledged by Hanushek's requirement that studies be published in a book or journal to assure a minimal quality check. The individual estimates that make up a study do not pass this quality hurdle in isolation: the combined weight of evidence in a study is evaluated to decide whether to publish it.

In view of the large differences between Hanushek's results and the results of the reanalysis undertaken here and in other meta-analyses, one should be reluctant to conclude that school resources are irrelevant to student outcomes. The strongest available evidence suggests a connection. In considering evidence on school resources and student achievement, it seems wise to raise the question asked by the Supreme Court of New Jersey in *Abbott v. Burke:* "[I]f these factors are *not* related to the quality of education, why are the richer districts willing to spend so much for them?"

Economics provides a useful framework for valuing the tradeoffs involved in increasing or decreasing class size. The calculations described in Section II, subject to the many caveats listed there, suggest that the economic benefits of further reductions in class size in grades K-3 are greater than the costs if a 4% real interest rate is used to discount benefits and costs to present values, and are about equal to the costs if a 6% real interest rate is used. With 1% per annum productivity growth and a 4% real discount rate, the "critical effect size" for the benefit of a reduction from 22 to 15 students to equal the costs is estimated to equal 0.10 standard deviations. This would be a natural hypothesis against which to test findings to judge their economic significance. Without knowing whether estimates are able to rule out the "critical effect size," it is difficult to assess the economic implications of the class size literature as a whole. The overall effect size from the STAR experiment, however, exceeds this critical effect size. Further, economic considerations suggest that greater gains might be available if resources were targeted toward those groups — minority and disadvantaged students — who appear to benefit the most from smaller classes.

Endnotes

1. See Krueger (1999a).

2. This quote is from Hanushek (1997, 148).

3. The word "studies" is in quotation marks because the unit of observation in Hanushek's work is not an entire study, but rather an individual estimate, of which several might be drawn from a single study. This point is discussed more fully below.

4. The distinction between studies and separate estimates is often blurred in the press. For example, an article in *Education Week* (April 12, 2000) on a class size reduction program in Wisconsin reported that Eric Hanushek "has examined more than 275 similar studies."

5. It is not uncommon for some of the estimates to be based on as few as 20 degrees of freedom (i.e., there are only 20 more observations than parameters to be identified), so the sampling errors can be very large.

6. The same data are used in the literature summaries in Hanushek (1996a, 1996b, and 1998).

7. Many of these studies reported more than one estimate, but only one estimate was selected because the separate estimates may not have been deemed sufficiently different in terms of sample or specification. Hanushek (1997) notes that as a general rule he tried to "reflect the estimates that are emphasized by the authors of the underlying papers."

8. It is unclear how Hanushek derived 24 estimates of unknown sign from this study, however, because no mention of the class size variable was made in connection to the equations for the reading scores.

9. In the Card and Krueger study, controlling for the income and education of parents leads to a slight increase in the effect of class size reductions on the rate of return to schooling.

10. I also follow the practice of using the terms "class size" and "pupil-teacher ratio" interchangeably. The difference is primarily a question of how one aggregates microdata.

11. The p-value was calculated assuming 59 independent Bernoulli trials, from the 59 studies used. If instead the number of independent Bernoulli trials was 277 — the number of estimates Hanushek extracted from the literature — the p-value in column 1 would be 0.32.

12. If the weights were selected to minimize the sampling variance of the combined estimate, the optimal weights would be the inverse of the sampling variances of the individual estimates (see Hedges and Olkin 1985).

13. For example, if a study was classified as having one estimate that was positive and significant and one that was positive and insignificant, these two categories would each be assigned a value of 50%, and the others would be assigned 0. If a study reported only one estimate, the corresponding category would be assigned 100% for that study.

14. The dependent variable in column 1, for example, is the percentage of a study's estimates that are positive and statistically significant; the independent variable is the number of estimates. Therefore, the intercept gives the expected percentage positive and significant if there are zero estimates. Adding the slope gives the expected percentage if exactly one estimate is extracted per study. Obviously, in a study with only one estimate, either zero or 100% of the estimates will be positive and significant. The expected percentage for one estimate per study can be interpreted as the probability that a study's single estimate will be positive and significant, or as the fraction of single-estimate studies that we expect to have positive and significant results. These expected percentages are reported in column 4 of Table 1-2.

15. Models estimated with this sample included eight explanatory variables and an intercept, so there were only 13 degrees of freedom. This is quite low, and would typically lead to very imprecise estimates.

16. This type of problem arises in many estimates that Hanushek uses because the underlying studies were not designed to study the effect of class size per se but some other feature of the education process. Maynard and Crawford, for example, were interested in the effect of exogenous shifts in family income (arising from income maintenance experiments) on children's academic outcomes, and the study provides persuasive results on this issue; class size and expenditures per pupil were just ancillary variables that the researchers held constant.

17. The assumption of optimal behavior by schools is supported by the theory of Tiebout sorting, in which it is an expected result of competition among municipalities. If, on the margin, parents chose where to live based on the schools, then one would expect schools to behave optimally. This, of course, stands in direct contradiction to the claims of Chubb and Moe (1990) and Finn (1991), who argue that schools do not optimize because their administrators are unaccountable and free of competition.

18. These results come from a multiple regression with the log of the wage as the dependent variable and indicators for the reading and math scores in the upper and lower quartiles as explanatory variables. Currie and Thomas also estimate separate regressions for men and women, controlling in these models for father's occupation, father's education, number of children and birth order, mother's

age, and birth weight. The wage gap between those who score in the top and bottom quartiles on the reading exam in these models is 13% for men and 18% for women, and on the math exam it is 17% for men and 9% for women. This suggests that only a modest part of the observed relationship between test scores and earnings results from differences in student background.

19. Folger and Parker (1990) tentatively conclude from the STAR experiment that proportionality is a reasonable assumption.

20 See *Digest of Education Statistics, 1998*, Table 169.

21. Students spent less than four years in a small class because half the students entered the experiment after the first year, and because some students moved to a new school or repeated a grade, causing them to return to regular size classes.

22. The figure is based on data from the March 1999 Current Population Survey. The sample consists of all civilian individuals with any work experience in 1998.

23. Formally, the average real wage for a worker who reaches age A in year t, denoted Y_t, is calculated by $Y_t = E_A(1+\gamma)t$, where E_A is the average earnings in Figure 1B for a worker of age A and γ is the rate of productivity growth.

24. This could be because students assigned to small classes lost ground as they progressed through the later grades, or because students initially assigned to regular classes caught up to the small-class students.

25. Because the costs are proportional to the teacher-pupil ratio, not to the number of students per teacher, the critical effect size for a one-student reduction varies depending on the initial class size.

26. Note, however, that Jencks and Phillips (1999) find that math test score gains between 10th and 12th grade have about the same impact on subsequent earnings as cross-sectional differences in scores of equivalent magnitude in 10th grade.

27. Betts (1996) similarly finds that the mean estimate in this literature is 1.04% higher earnings for 10% greater spending.

28. This was calculated by $0.010 = 0.08*0.20*2*0.1/(7/22)$. One difficulty in comparing these two literatures, however, is that it is unclear how long class size is reduced in the observational studies on earnings. In some studies, the pupil-teacher ratio during a student's entire elementary and secondary school career is used, while in others just one year's data are used.

References

Angrist, Joshua, and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Children's Academic Achievement." *Quarterly Journal of Economics* 114(2): 533-75.

Betts, Julian R. 1996. "Is There a Link Between School Inputs and Earnings?" In Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington D.C.: Brookings Institution, pp. 141-91.

Burkhead, J. 1967. *Input-Output in Large City High Schools.* Syracuse, N.Y.: Syracuse University Press.

Card, David. 1999. "The Causal Effect of Schooling on Earnings." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics,* Amsterdam: North Holland. Forthcoming.

Card, David, and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100 (February): 1-40.

Card, David, and Alan B. Krueger. 1996. "Labor Market Effects of School Quality: Theory and Evidence." In Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington D.C.: Brookings Institution, pp. 97-140.

Cawley, John, Karen Conneely, James Heckman, and Edward Vytlacil. 1996. "Measuring the Effects of Cognitive Ability." Working Paper No. 5645. Cambridge, Mass.: National Bureau of Economic Research.

Chubb, John E., and Terry M. Moe. 1990. *Politics, Markets and America's Schools.* Washington, D.C.: Brookings Institution.

Currie, Janet, and Duncan Thomas. 1999. "Early Test Scores, Socioeconomic Status and Future Outcomes." Working Paper No. 6943. Cambridge, Mass.: National Bureau of Economic Research.

Finn, Chester E. 1991. *We Must Take Charge,* New York: Free Press.

Finn, Jeremy D., and Charles M. Achilles. 1990. "Answers and Questions About Class Size: A Statewide Experiment." *American Educational Research Journal* 27 (Fall): 557-77.

Folger, John, and Jim Parker. 1990. "The Cost-Effectiveness of Adding Aides or Reducing Class Size." Vanderbilt University, mimeo.

Hanushek, Eric A. 1981. "Throwing Money at Schools." *Journal of Policy Analysis and Management* 1(1): 19-41.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24 (September): 1141-77.

Hanushek, Eric A. 1989. "Expenditures, Efficiency, and Equity in Education: The Federal Government's Role." *American Economic Review* 79(2): 46-51.

Hanushek, Eric A. 1996a. "A More Complete Picture of School Resource Policies." *Review of Educational Research* 66: 397-409.

Hanushek, Eric A. 1996b. "School Resources and Student Performance." In Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington D.C.: Brookings Institution, pp. 43-73.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19(2): 141-64.

Hanushek, Eric A. 1998. "The Evidence on Class Size." Occasional Paper Number 98-1. Rochester, N.Y.: W. Allen Wallis Institute of Political Economy, University of Rochester.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis.* Orlando, Fla.: Academic Press.

Hedges, Larry V., Richard Laine, and Rob Greenwald. 1994. "Does Money Matter? A Meta- Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Education Researcher* 23(3): 5-14.

Jencks, Christopher S., and M. Brown. 1975. "Effects of High Schools on their Students." *Harvard Educational Review* 45(3): 273-324.

Jencks, Christopher S., and Meredith Phillips. 1999. "Aptitude or Achievement: Why Do Test Scores Predict Educational Attainment and Earnings?" In Susan Mayer and Paul Peterson, eds., *Learning and Earning: How Schools Matter.* Washington, D.C.: Brookings Institution Press. Forthcoming.

Kiesling, H. J. 1967. "Measuring a Local Government Service: A Study of School Districts in New York State." *Review of Economics and Statistics* 49: 356-67.

Krueger, Alan B. 1999a. "Measuring Labor's Share." *American Economic Review* 89(2): 45-51.

Krueger, Alan B. 1999b. "Experimental Estimates of Educational Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B., and Diane Whitmore. 1999. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence From Project STAR." Working Paper No. 427. Princeton, N.J.: Princeton Industrial Relations Section.

Lazear, Edward P. 1999. "Educational Production." Working Paper No. 7349. Cambridge, Mass.: National Bureau of Economic Research.

Link, Charles R., and James G. Mulligan. 1986. "The Merits of a Longer School Day." *Economics of Education Review* 5(4): 373-81.

Link, Charles R., and James G. Mulligan. 1991. "Classmates' Effects on Black Student Achievement in Public School Classrooms." *Economics of Education Review* 10(4): 297-310.

Maynard, Rebecca, and D. Crawford. 1976. "School Performance." *Rural Income Maintenance Experiment: Final Report*. Madison: University of Wisconsin.

Mosteller, Frederick. 1995. "The Tennessee Study of Class Size in the Early School Grades." *The Future of Children: Critical Issues for Children and Youths* 5 (Summer/Fall): 113-27.

Murnane, Richard, John Willet, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77: 251-66.

Neal, Derek, and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differentials." *Journal of Political Economy* 104 (October): 869-95.

Nye, Barbara, Jayne Zaharias, B.D. Fulton, et al. 1994. "The Lasting Benefits Study: A Continuing Analysis of the Effect of Small Class Size in Kindergarten Through Third Grade on Student Achievement Test Scores in Subsequent Grade Levels." Seventh grade technical report. Nashville: Center of Excellence for Research in Basic Skills, Tennessee State University.

Summers, Anita and B. Wolfe. 1977. "Do Schools Make A Difference?" *American Economic Review*, 67 (4), pp. 649-52.

PART II: EVIDENCE, POLITICS, AND THE CLASS SIZE DEBATE

Eric A. Hanushek

With the suddenness of a summer storm, politics thrust the issue of class size policy onto the national agenda. Before the political popularity to voters of reductions in class size became known, most educational researchers and policy makers had discarded such policies as both too expensive and generally ineffective, leaving only teachers unions and others with clear vested interests in the policies to support such ideas. When the political appeal of class size reductions became known — largely through the reactions to the 1996 California policies — there was a scramble to backfill evidence supporting such policies. In this current environment, the evidence about the effectiveness of class size reduction has been thoroughly spun in the political debate in order to match the preconceived policy proposals, making it difficult to conclude that the debate has been guided very much by the evidence.

This political backdrop is necessary to understand the significance of Alan Krueger's reanalysis of the existing evidence on class size (Krueger 2000). He focuses attention directly on the scientific evidence and its implications for policy, thus attempting to move the policy debate away from pure politics and toward a better basis for decision making. While he offers no new evidence on the effects of class size on student performance, he contributes two different analyses that point toward a more aggressive policy of class size reduction: a massaging of the econometric evidence on effectiveness of class size reduction and of overall spending and a proposed demonstration that small outcome effects are still worthwhile. Upon careful inspection, however, neither is convincing. Nonetheless, policy makers should not ignore the emphasis on the importance of a solid evidentiary base.

Because supporters of class size reductions are likely to be attracted to his defense of such policies, it is important to understand the nature and substance of his analysis. First, his discussion omits mention of the long history and dismal results of class size policies. Second, his analysis of the existing econometric evidence derives its results from giving excessive weight to low-quality and biased estimates. Third, the discussion of the Tennessee STAR (Student/Teacher Achievement Ratio) experiment does not make clear its limited evidence for any broad reductions and fails to indicate the uncertainty surrounding the results and their policy implications. Finally, the calculation of benefit-cost relationships takes a very narrow view of potential policies and requires a number of heroic assumptions. This set of comments discusses each of these in turn.

The issue of course is not whether there exists any evidence that class size reduction *ever* matters. Surely class size reductions are beneficial in specific circumstances — for specific groups of students, subject matters, and teachers. The policy debates, driven by the politics of the situation, do not, however, attempt to identify any such specific situations but instead advocate broad reductions in class sizes across all schools, subjects, and often grades. The missing elements are three. First, nothing in the current decision process encourages targeting class size reductions to situations where they are effective. Second, class size reductions necessarily involve hiring more teachers, and teacher quality is much more important than class size in affecting student outcomes. Third, class size reduction is very expensive, and little or no consideration is given to alternative and more productive uses of those resources.

Similarly, while some have characterized my past research as indicating that "money makes no difference," this summary is inaccurate and misleading. My research and that of others shows that there are large differences among teachers and schools — differences that should be in my opinion the focus of aggressive public policy. At the same time, the organization of schools and the attendant incentives to improve

student performance have been shown to distort the gains that could potentially come from added resources to schools. While some schools may use added resources to improve student outcomes, others will not. Moreover, we do not have the ability to predict which schools and which uses of additional funds will be effective. Therefore, the correct summary is "just providing more resources — whether in the form of reduced class sizes or in other forms — is unlikely to lead to higher student achievement as long as future actions of schools are consistent with their past choices and behavior."

The appeal of class size reduction is that it offers the hope of improving schools while requiring no change in the existing structure. Politicians can take credit for pursuing identifiable policies aimed at improving student outcomes. Teachers and other school personnel see added resources coming into schools without pressures to take responsibility for student performance and see these policies increasing the demand for teachers. The missing element is any reasonable expectation that these policies will significantly improve student achievement.

I. The history of class size reduction

Perhaps the most astounding part of the current debates on class size reduction is the almost complete disregard for the history of such policies. Pupil-teacher ratios fell dramatically throughout the 20th century.[1] **Table 2-1** shows that pupil-teacher ratios fell by a third between 1960 and 1995 — exceeding the magnitude of policy changes that most people are talking about today. With such substantial changes, one would expect to see their effect in student performance. Yet it is impossible to detect any overall beneficial effects that are related to these sustained increases in teacher intensity.

The longest general data series on student performance, albeit imperfect, is the Scholastic Aptitude Test (SAT). **Figure 2A** displays the relationship between pupil-teacher ratios and SAT scores. While there is a relationship between the two, it goes in the opposite direction expected: reductions in pupil-teacher ratios are accompanied by falls in the SAT, even when appropriately lagged for the history of schooling experience for each cohort of students. Because the SAT is a voluntary test taken by a select population, a portion of the fall undoubtedly reflects changes in the test-taking population instead of real declines in aggregate student performance, but there is general consensus that real declines also occurred (Congressional Budget Office 1986).

A better indicator of performance is the National Assessment of Educational Progress (NAEP). While tracking a representative sample of students, scores are only available since the early 1970s (after a period of substantial decline as measured by the SAT). **Figure 2B** plots NAEP scores for 17-year-olds.[2] Math and reading show flat performance from earliest testing through 1996, while the comparable science and writing scores have declined significantly.[3] Thus, the consistent picture from available evidence is that the falling pupil-teacher ratios (and commensurately increasing real spending per pupil) have not had a discernible effect on student achievement.

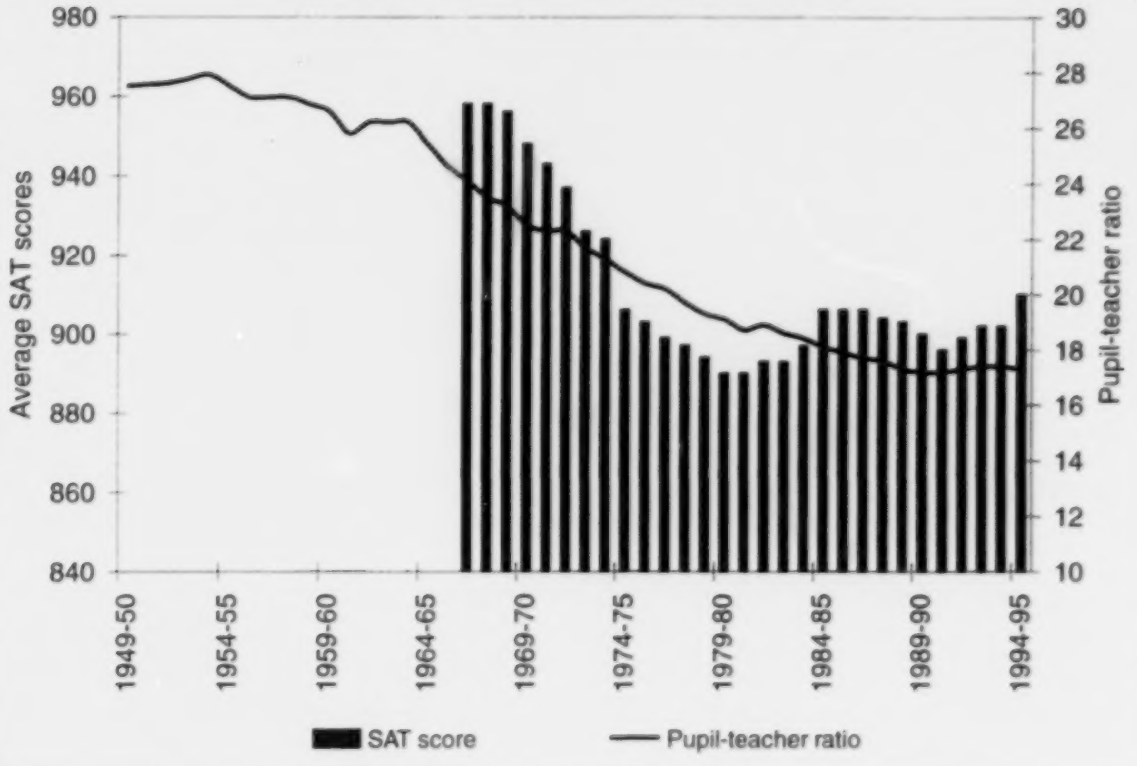
While it is generally difficult to infer causation from aggregate trends, these data provide a strong *prima facie* case that the policies being discussed today will not have the significant outcomes that are advertised. The complication with interpreting these trend data is that other factors might work to offset an underlying beneficial effect. On this, the available evidence does not indicate that the pattern of test scores simply reflects changing student characteristics. Child poverty and the incidence of children in single-parent families — factors that would be expected to depress achievement — have risen. At the same time, the increases in parental education and the fall in family sizes would be expected to produce improvements in student performance. Netting out

TABLE 2-1
Pupil-teacher ratio and real spending, 1960-95

	1960	1970	1980	1990	1995
Pupil-teacher ratio	25.8	22.3	18.7	17.2	17.3
Current expenditure per pupil (1996/97 $)	$2,122	$3,645	$4,589	$6,239	$6,434

FIGURE 2A

Pupil-teacher ratios and student performance

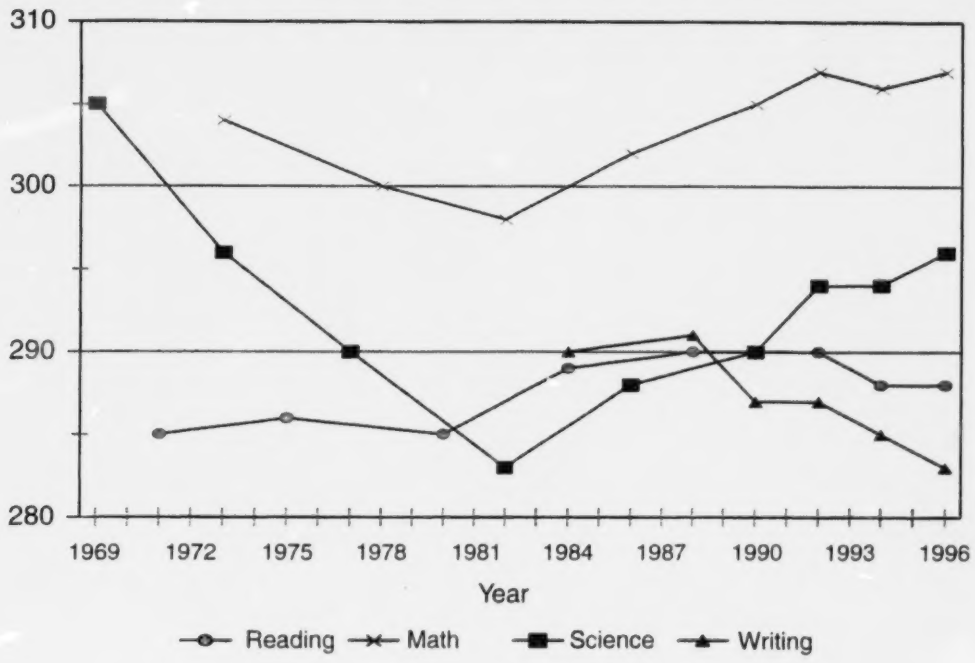


these effects is difficult to do with any precision, but the existing analysis suggests little aggregate effect from the changing student backgrounds, and possibly a small net improvement.[4]

Table 2-1 also shows the significant increases in expenditure per pupil that have occurred over this period. A significant part of the increase in expenditure can be directly attributable to declines in the pupil-teacher ratio (Hanushek and Rivkin 1997), but other "improvements" such as having a more experienced and educated teacher force also contribute. Again, however, a comparison of student performance with the increases in inflation-adjusted expenditures of over 75% between 1970 and 1995 gives no reason to believe that more of the past resource policies will be successful.

FIGURE 2B

National assessment of educational progress, 17-year-olds



If past declines in class size have had no discernible effect on student outcomes, why should we believe that future declines would yield any different results?

II. Econometric evidence

Krueger (2000) concentrates most of his attention on the existing econometric evidence. While worrying about important issues, the analysis actually involves a set of calculations that places the heaviest weight on lower-quality estimates. By doing so, he is able to suggest that the overall conclusions about class size policies should change. If, however, more weight is placed on higher-quality estimates, the overall conclusion about a lack of clear relationship between class size and student performance is strengthened.

The starting point of Krueger's work is my prior tabulations of the estimated relationship between teacher-pupil ratios on student performance, as reproduced in **Table 2-2**.[5] The 277 separate estimates of the class size relationship are found in 59 publications, representing all of the available analyses through 1994. (Issues about the underlying data raised by Krueger (2000) do not change any of the results, and a discussion of them is included in the appendix to these comments).

Among the statistically significant estimates — the ones for which we are reasonably confident that there is truly a relationship — 14% indicate that raising the teacher-pupil ratio would have the "expected"

TABLE 2-2
Percentage distribution of estimated effect of teacher-pupil ratio and spending on student performance

Resource	Number of estimates	Statistically significant		Statistically insignificant		
		Positive	Negative	Positive	Negative	Unknown sign
Teacher-pupil ratio	277	14%	14%	27%	25%	20%
Expenditure per pupil	163	27	7	34	19	13

Source: Hanushek (1997), as corrected (see text).

TABLE 2-3
Sample sizes for estimated effect of teacher-pupil ratio by number of estimates per publication

Number of estimates per publication	Number of estimates (publications)	Sample size			
		Median	Average	Minimum	Maximum
1	17 (17)	272	1,310	48	14,882
2-3	28 (13)	649	1,094	47	5,000
4-7	109 (20)	512	2,651[a]	38	18,684[b]
8-24	123 (9)	266	1,308	22	10,871
1-24	277 (59)	385	1,815[c]	22	18,684[d]

Note: Because of ambiguity about the precise estimation in Harnisch (1987), an alternative calculation of sample sizes that represent schools rather than individuals is given: a. 1,996; b. 16,456; c. 1,557; d. 16,456.

positive relationship while an equal percentage indicate just the opposite. The statistically insignificant estimates — those for which we have less confidence that they indicate any real relationship — are almost evenly split between beneficial and adverse effects. Thus, the overall evidence provides little reason to believe that a general policy of class size reduction would improve student performance.

Krueger questions these conclusions by arguing that individual publications that include more separate estimates of the impact of class size on performance are lower in quality than those publications that include fewer estimates.[6] His hypothesis is that publications including more estimates will involve splitting the underlying samples of student outcomes, say by race or grade level. Statistical theory indicates that, other things being equal, smaller samples will yield less precise estimates than larger samples. He then jumps to the logically incorrect conclusion that publications with more individual estimates will tend to have fewer observations and thus will tend to produce statistically insignificant results when compared to those publications with fewer separate estimates.

There is no clear relationship between the sample sizes underlying individual estimates and the number of estimates in each publication. **Table 2-3** shows the distribution of sample sizes for the 277 estimates of the effect of teacher-pupil ratios from Table 2-2. While highly variable, publications with the fewest estimates do not

systematically have the largest sample sizes. The simple correlation of sample sizes and number of articles in the underlying publications is slightly positive (0.03), although insignificantly different from zero.[7]

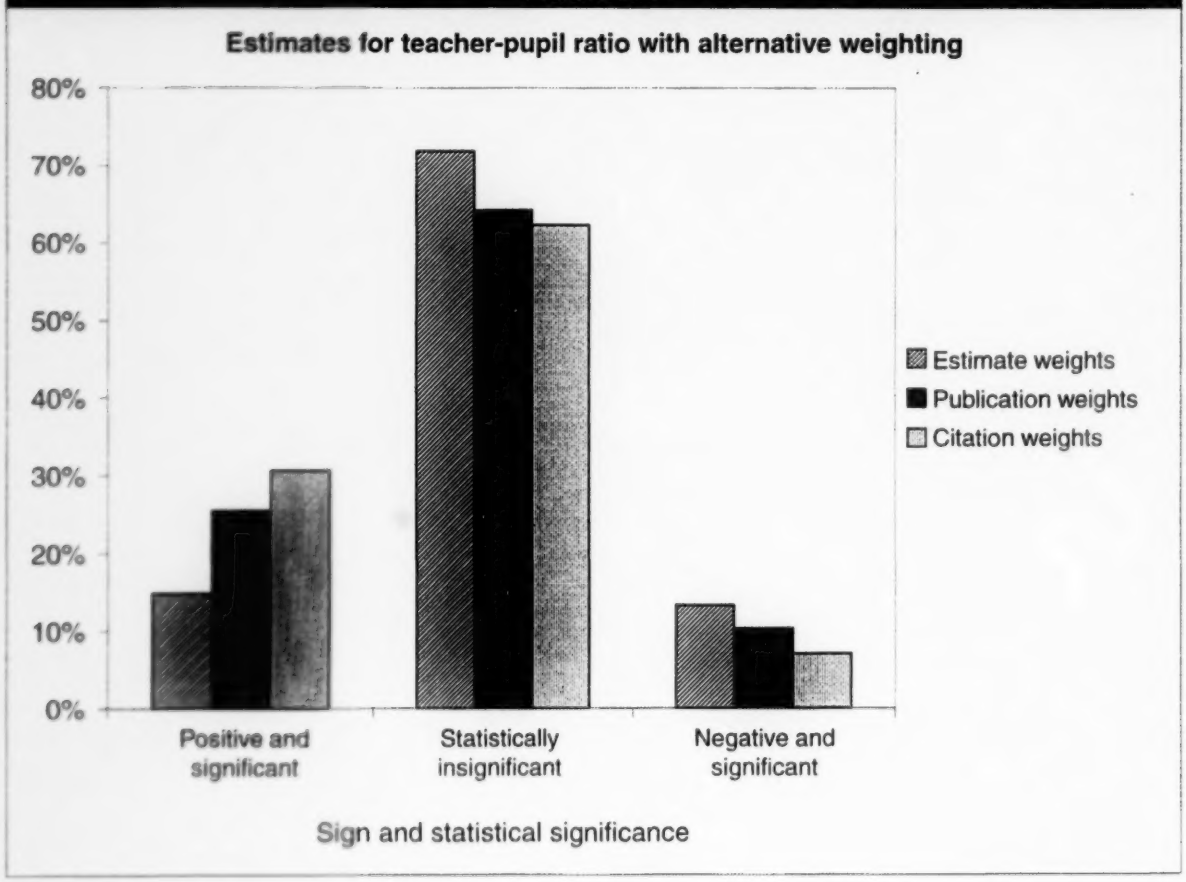
Before considering the precise nature of Krueger's re-analysis, it is useful to understand better the structure of the underlying estimates and publications. The explanation for varying numbers of estimates across individual publications is best made in terms of the provision of logically distinct aspects of the achievement process. For example, few people argue that the effects of class size reduction are constant across all students, grades, and subject matter. Therefore, when the data permit, researchers will typically estimate separate relationships for different students, different outcomes, and different grades. In fact, the analysis of the Tennessee class size experiment in Krueger (1999) divides the estimates by race and economic status, because Krueger himself thought it was plausible that class size has varying impacts — something that he finds and that he argues is important for policy. He further demonstrates varying effects by grade level. If there are different effects for different subsamples of students, providing a single estimate across the subsamples, as advocated by Krueger (2000) and described below, is incorrect from a statistical point of view and would lead to biased results.

Even if class size differences have similar effects across students, districts, and outcomes, it is often impossible to combine the separate samples used for obtaining the individual estimates. For example, the publication by Burkhead et al. (1967) that Krueger holds up as an example of multiple estimates for small samples presents a series of estimates for high school performance in different cities where outcomes are measured by entirely different instruments. There is no way in which these can be aggregated into a single estimate of the effect of class size. Of the 59 publications from Table 2-2 that include estimates of the effects of teacher-pupil ratio, 34 include two or more separate test measures of outcomes (e.g., reading and math), and 15 of these further include two or more separate non-test measures (e.g., college continuation, dropouts, or the like). For 14 of the 59 publications, the separate estimates of pupil-teacher effects within individual publications include students separated by more than three grade levels, implying not only different achievement tests but also the possibility of varying effects across grades. No general procedure exists for aggregating these separate effects in a single econometric estimate.

Thus, while Krueger suggests that the publication of multiple estimates is largely whimsical and misguided, the reality is that there are generally sound econometric reasons behind many of these decisions. The typical publication with several estimates actually provides more evidence than would be the case if only one estimate per publication were reported.

Krueger's hypothesis, however, is that an estimate in publications with more than one estimate provides poorer information than an estimate from a single-estimate publication. His analytical approach involves adding up the underlying estimates in alternative ways — effectively giving increased weight to some estimates and decreased weight to others. Specifically, he calculates the proportion of estimates within each publication that fits into the outcome categories (columns) in Table 2-2 and adds them up across the 59 separate publications, i.e., weighting by individual publications instead of individual estimates of the effect of class size on student performance. Surprisingly, this procedure leads to stronger support for the existence of positive effects from class size reduction, even though the simple statistical theory outlined by Krueger suggests that only the confidence in the estimates and not the direction of the relationship should be affected. The evidence based on the estimates in Table 2-2 indicates an essentially identical chance of finding increased teacher-pupil ratios to be beneficial as a chance of being harmful; i.e., no systematic relationship between class size and student outcomes. When re-weighted, however, Krueger finds beneficial effects to be noticeably more likely.

FIGURE 2C



Note, however, that still only 25% of the time would there be much confidence that there is a relationship between teacher-pupil ratios and achievement as indicated by their being a statistically significant and positive estimate. To reach his conclusions of different overall results, Krueger tends to emphasize the proportion of estimates that are positive (beneficial) versus negative (detrimental). This summary has a major problem. The equal weighting of statistically significant estimates (those more precisely estimated) and statistically insignificant estimates (less precisely estimated) seems to violate the basic premise of his re-weighting.[8] A more accurate picture of the impact of his weighting is seen in **Figure 2C**, which graphs the proportion of results that are statistically significant (positive or negative) and that are statistically insignificant. His re-weighting produces a somewhat higher proportion of positive and statistically significant results, but it does not reverse the overall picture of little reason to expect much if any impact from reducing class size.[9]

To deal with the apparent anomaly of finding different results when re-weighted, Krueger introduces a "theory of refereeing" for scholarly publications. He suggests that, whenever an author finds results that are statistically insignificant or that have the wrong sign, referees will insist that the author re-do the estimates by disaggregating them — in effect producing more of the insignificant or wrong-signed estimates.

While Krueger provides no evidence for his theory of refereeing, many — including Krueger himself

— have argued just the opposite about the publication process. Specifically, there is a well-known publication bias toward having too many statistically significant estimates in articles that get published. Articles with insignificant estimates or incorrect signs simply do not get published with the same frequency as articles containing significant estimates of the expected sign (Hedges 1990). Krueger's own argument in discussing the literature on the minimum wages is "reviewers and editors have a natural proclivity to look favorably on studies that report statistically significant results" (Card and Krueger 1995, 186).

Krueger is correct about the importance of quality of the estimates in formulating overall conclusions, and consideration of quality provides a much more natural and persuasive explanation for his altered results than does his theory of refereeing. The basic tabulation of results produced in Table 2-2 provided information on all available estimates of the effects of class size and of spending. The complete data are displayed not as an endorsement of uniform high quality but as a base case where there can be no possibility that selection of specific estimates and publications drives the results. At the same time, the underlying analyses clearly differ in quality, and — as discussed in Hanushek (1997) — these differences have the potential for biasing the results of the estimation. Two elements of quality are particularly important. First, education policy in the United States is made primarily by the separate 50 states, and the variations in spending, regulations, graduation requirements, testing, labor laws, and teacher certification and hiring policies are large. These important differences — which are also the locus of most current policy debates — imply that any analyses of student performance across states must include descriptions of the policy environment of schools or else they will be subject to standard statistical bias problems; i.e., they will tend to obtain estimates that are systematically different from reality. Second, education is a cumulative process going across time and grades, but a majority of estimates consider only the current resources available to students in a given grade. For example, when looking at performance at the end of secondary schooling, many analyses rely on just the current teachers and school resources and ignore the dozen or more prior years of inputs. Obviously, current school inputs will tend to be a very imperfect measure of the resources that went into producing ending achievement.

While judgments about study quality generally have a subjective element, it is possible to make an initial cut based on the occurrence of these two problems. We begin with the issue of not measuring the state policy environment. If, as most people believe, states vary in important aspects of education policy and school operations, ignoring this in the econometric estimation will generally lead to biased estimates of the effect of teacher-pupil ratios or other resources.[10] The key is separating the effects of teacher-pupil ratios from other attributes of schools and families, and this generally cannot be done accurately if the other factors are not explicitly considered. Whether the estimates tend to find too large or too small effect of teacher-pupil ratios depends on the correlation of the omitted state regulatory and finance factors and class size (or spending).

The existing estimates contained in Table 2-2 can be used to identify the importance of biases caused by omitting consideration of differences in the state policy environment for schools. Specifically, an analysis that looks at schools entirely contained within a single state will observe a policy environment that is largely constant for all schools — and thus the econometric estimates that compare schooling entirely within a single state will not be biased. On the other hand, an analysis that considers schools in multiple states will produce biased results whenever important state differences in policy are correlated with differences across states in pupil-teacher ratios or overall resources. Moreover, the statistical bias will be largest for investigations relying on aggregate state data as opposed to observations at the classroom or school level.[11]

Thus, one clear measure of study quality is that it relies upon data entirely within a single state. For those using multistate data, estimates derived from the most aggregated data will be of lower quality than those relying on observed resources and outcomes at the classroom or school level.

TABLE 2-4
Percentage distribution of estimated effect of teacher-pupil ratio and expenditure per pupil by state sampling scheme and aggregation

Level of aggregation of resources	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
A. Teacher-pupil ratio				
Total	277	14%	14%	72%
Single state samples[a]	157	11	18	71
Multiple state samples[b]	120	18	8	74
Disaggregated within states[c]	109	14	8	78
State-level aggregation[d]	11	64	0	36
B. Expenditure per pupil				
Total	163	27%	7%	66%
Single state samples[a]	89	20	11	69
Multiple state samples[b]	74	35	1	64
Disaggregated within states[c]	46	17	0	83
State-level aggregation[d]	28	64	4	32

a. Estimates from samples drawn within single states.
b. Estimates from samples drawn across multiple states.
c. Resource measures at level of classroom, school, district, or county, allowing for variation within each state.
d. Resource measures aggregated to state level with no variation within each state.

Table 2-4 provides a tabulation of the prior econometric results that is designed to illuminate the problem of ignoring the large differences in school organization and policy across states. The prior tabulation of all estimates shows that those with significant negative estimates evenly balance the percentage indicating teacher-pupil ratios with significant positive estimates. But Table 2-4 shows that this is not true for estimates relying upon samples drawn entirely within a single state, where the overall policy environment is constant and thus where any bias from omitting overall state policies is eliminated. For single state analyses, the statistically significant effects are disproportionately negative (18% negative versus 11% positive). Yet, when the samples are drawn across states, the relative proportion positive and statistically significant rises. For those aggregated to the state level, almost two-thirds of the estimates are positive and statistically significant. The pattern of results also holds for estimates of the effects of expenditure differences (where positive and statistically significant estimates are most likely to come from investigations involving both multiple states and data aggregated to the state level).[12] Again, the vast majority of estimates are statistically insignificant or negative in sign except for those employing aggregated state-level data and neglecting differences in state policy environments. This pattern of results is consistent with expectations from considering specification biases when favorable state policies tend to be positively correlated with resource usage, i.e., when states with the best overall education policies also tend to have larger teacher-pupil ratios and higher spending.

The second problem is that the cumulative nature of the educational process means that relating the level of performance at any point in time just to the current resources is likely to be misleading. The mismeasurement is strongest for any children who changed schools over their career (a sizable majority in the

TABLE 2-5
Percentage distribution of estimates of teacher-pupil ratio on student performance, based on value-added models of individual student performance

	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
All	798	11%	9%	80%
Estimates for single state samples	24	4%	17%	79%

U.S.) but also holds for students who do not move because of variations over time in school and family factors. While there is no general theoretical prediction about the biases that arise from such mismeasurement, its importance can be understood by concentrating on estimates that do not suffer from the problem. The standard econometric approach for dealing with this is the estimation of value-added models where the statistical estimation is restricted to the growth of achievement over a limited period of time (where the flow of resources is also observed). By concentrating on achievement gains over, say, a single grade, it is possible to control for initial achievement differences (which will have been determined by earlier but generally unobserved resources and other educational inputs).

Table 2-5 displays the results of teacher-pupil ratio estimates that consider value-added models for individual students. The top panel shows all such results, while the bottom panel follows the earlier approach of concentrating just on estimates within an individual state. With the most refined investigation of quality in the bottom panel, the number of estimates gets quite small and selective. In these, however, there is essentially no support for a conclusion that higher teacher-pupil ratios improve student performance. Only one of the available 24 estimates (4%) shows a positive and statistically significant relationship with student outcomes, while 17% find a negative and statistically significant relationship.

As noted previously, teacher-pupil ratios and class size are not the same measure, even though they tend to move together. The general estimation in Table 2-2 makes no distinction between the two measures. In the case of estimation at the individual classroom (the focus of Table 2-5), however, the teacher-pupil ratio is essentially the same as class size. Thus, those measurement issues cannot distort these results.

This direct analysis of study quality shows why Krueger gets different effects from weighting results by publication instead of by individual estimates. From Table 2-2, 17 of the 59 publications (29%) contained a single estimate of the effect of the teacher-pupil ratio — but these estimates are only 6% of the 277 total available estimates. Krueger wants to increase the weight on these 17 estimates (publications) and commensurately decrease the weight on the remaining 260 estimates. Note, however, that over 40% of the single-estimate publications use state aggregate data, compared to only 4% of all estimates.[13] Relatedly, the single-estimate publications are more likely to employ multistate estimates (which consistently ignore any systematic differences in state policies) than the publications with two or more estimates. Weighting by publications rather than separate estimates heavily weights low-quality estimates.

The implications are easy to see within the context of the two publications that Krueger himself contributes (Card and Krueger 1992a, 1992b). Each of these state-level analyses contributes one positive, statistically significant estimate of the effect of teacher-pupil ratios. Weighting by all of the available estimates, these

estimates represent 0.7% of the available estimates, but, weighting by publications as Krueger desires, they represent 3.4% of the results. Krueger (2000) goes on to say that Card and Krueger (1992a) "presented scores of estimates for 1970 and 1980 Census samples sometimes exceeding one million observations. Nonetheless, Hanushek extracted only one estimate from this study because only one specification included family background information." This statement is quite misleading, however. While the underlying Census data on earnings included over a million observations, the relevant estimate of the effects of class size in Card and Krueger (1992a) relies on *just 147 state aggregate data points* representing different time periods of schooling.[14]

Krueger's statement also implies that requiring information on family backgrounds is some sort of irrelevant technicality. There are, however, very important econometric reasons for insisting on the inclusion of family background as a minimal quality requirement. It is well known that family background has a powerful effect on student performance (see, for example, Coleman et al. (1966) or Hanushek (1992)). If this factor is omitted from the statistical analysis, the estimates of pupil-teacher ratios can no longer be interpreted as the effect that class size might have on student performance. These estimates will be biased if there is any correlation across states between family backgrounds, such as income and education, and the average teacher-pupil ratio in the state. Considering estimates that do not take varying family backgrounds into account is a very significant quality problem, because estimates of the effect of variations in pupil-teacher ratios will then reflect family background and will appear to be important even when pupil-teacher ratios have *no* impact on student performance. Such an omission almost certainly leads to larger distortions than does considering estimates that do not consider the state policy environment.

In fact, Card and Krueger (1992b) was mistakenly included in the tabulations. Discussions with Krueger about the coding of the full set of estimates made it clear that this publication failed to take any aspect of family background into account, so it cannot adequately distinguish school effects from family effects on learning. The concern, as discussed above, is that family background and pupil-teacher ratios tend to be correlated, so that — if family background is omitted from the analysis — the estimated effect of the pupil-teacher ratio will not indicate the causal impact of differing pupil-teacher ratios but instead will just be a proxy for family background. While the analysis in Card and Krueger (1992b) stratifies by race or allows for a difference in the overall level of performance by race (i.e., an intercept dummy variable), the estimated effects for pupil-teacher ratio come from variations across states and over time in class size, when race is not observed to vary.[15] Similarly, Card and Krueger (1992a) estimates models just for white males, and Krueger asserts that this is the same as the stratification by race in Link and Mulligan (1991). Link and Mulligan (1991), however, include the racial composition of classrooms in their analysis, thus allowing them to sort out family background differences of classrooms from class size differences in a way that simple stratification does not. Given their analysis, there is no way to conclude that the Card and Krueger estimates of the pupil-teacher ratio are anything more than simply an indication of family background differences on student outcomes.

Finally, the Card and Krueger (1992a) analysis suffers not only from the biases of aggregate, cross-state analysis discussed previously but also from another set of fundamental shortcomings. They estimate state differences in the value of additional years of schooling according to 1980 Census information on labor market earnings and the state where workers were born (assumed to proxy where they were educated). They then relate the estimated value of a year of schooling to characteristics of the average school resources in the state in the years when a worker of a given age would have attended school. As critiques by Speakman and Welch (1995) and Heckman, Layne-Farrar, and Todd (1996a, 1996b) show, their estimates are very sensitive to the specific estimation procedure. Moreover, the state earnings differences cannot be interpreted in terms of school quality

differences in the way that Card and Krueger interpret them. In order to obtain their estimates of school quality, Card and Krueger (1992a) must assume that the migration of people across states is random and not based on differential earnings opportunities. Heckman, Layne-Farrar, and Todd (1996a, 1996b) show that there is selective migration and that this fundamental requirement for their interpretation is untrue.[16] Statistical shortcomings such as these can be identified in other estimates, but this example illustrates why the mechanical re-weighting proposed by Krueger (2000) can in fact push the results in a biased direction.

The alternative weighting methods of Krueger (2000) provide no better adjustments for anything that looks like quality of estimates. The two Card and Krueger articles are heavily cited in other articles, so that their combined weight increases to *17% of the total evidence* on a citation basis. But again this new weighting does not give an accurate estimate of the quality of the underlying estimates.[17] Similarly, the "selection-adjusted" weights place more emphasis on a positive and significant estimate if there was an estimated higher probability of getting a positive and significant estimate in an article (based solely on the number of estimates within each publication). The rationale behind this novel approach is entirely unclear and has no statistical basis.

In sum, Krueger's reanalysis of the econometric evidence achieves different results by emphasizing low-quality estimates. The low-quality estimates are demonstrably biased toward finding significant positive effects of class size reduction and of added spending. Remarkably, even when re-weighted, the support of overall class size reduction policies remains weak. Most of the estimates, no matter how tabulated, are not statistically different from zero at conventional levels.

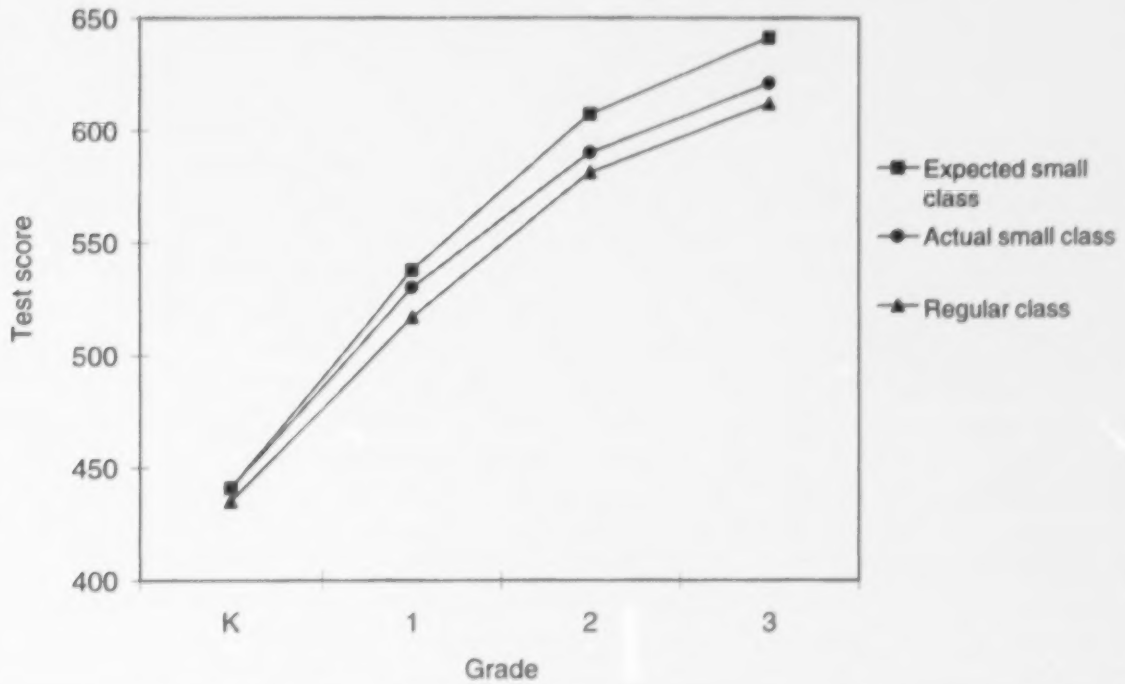
III. The Tennessee Class Size Experiment (Project STAR)

A different form of evidence — that from random assignment experiments — has recently been widely circulated in the debates about class size reduction. Following the example of medicine, one large-scale experimental investigation in Tennessee in the mid-1980s (Project STAR) pursued the effectiveness of class size reductions. Random-assignment experiments in principle have considerable appeal. The underlying idea is that we can obtain valid evidence about the impact of a given well-defined treatment by randomly assigning subjects to treatment and control groups. This eliminates the possible contaminating effects of other factors and permits conceptually cleaner analysis of the outcomes of interest across these groups. The validity of any particular experiment nonetheless depends crucially on the implementation of the experiment. On this score, considerable uncertainty about the results is introduced. But, ignoring any issues of uncertainty, the estimated impacts of large class size reductions are small and have limited application to the current policy proposals.

Project STAR was designed to begin with kindergarten students and to follow them for four years (Word et al. 1990). Three treatments were initially included: small classes (13-17 students); regular classes (22-25 students); and regular classes (22-25 students) with a teacher's aide. Schools were solicited for participation, with the stipulation that any school participating must be large enough to have at least one class in each treatment group. The initial sample included 6,324 kindergarten students. These were split between 1,900 in small classes and 4,424 in regular classes. (After the first year, the two separate regular class treatments were effectively combined, because there were no perceived differences in student performance).[18] The initial sample included 79 schools, although this subsequently fell to 75. The initial 326 teachers grew slightly to reflect the increased sample size in subsequent grades, although of course most teachers are new to the experiment at each new grade.

FIGURE 2D

Expected vs. actual STAR results, Stanford Achievement Test, reading



The results of the Project STAR experiment have been widely publicized. The simplest summary is that students in small classes performed significantly better than those in regular classes or regular classes with aides in kindergarten and that the achievement advantage of small classes remained constant through the third grade.[19]

This summary reflects the typical reporting, focusing on the differences in performance at each grade and concluding that small classes are better than large (e.g., Finn and Achilles 1990; Mosteller 1995). But it ignores the fact that one would expect the differences in performance to become wider through the grades because they continue to get more resources (smaller classes) and these resources should, according to the hypothesis, keep producing a growing advantage. **Figure 2D** shows the difference in reading performance in small classes that was observed across grades in Project STAR. (The results for math performance are virtually identical in size and pattern). It also shows how the observed outcomes diverge from what would be expected if the impact in kindergarten were also obtained in later grades. As Krueger (1999) demonstrates, the small class advantage is almost exclusively obtained in the first year of being in a small class — suggesting that the advantages of small classes are not general across all grades.

The gains in performance from the experimental reduction in class size were relatively small (less than 0.2 standard deviations of test performance), especially in the context of the magnitude of the class size reduction (around eight students per class). Thus, even if Project STAR is taken at face value, it has relatively limited policy implications.

While the experimental approach has great appeal, the actual implementation in the case of Project STAR introduces uncertainty into these estimates (Hanushek 1999b). The uncertainty arises fundamentally from questions about the quality of the randomization in the experiment. In each year of the experiment, there was sizable attrition from the prior year's treatment groups, and these students were replaced with new students. Of the initial experimental group starting in kindergarten, 48% remained in the experiment for the entire four years.[20] No information, such as pretest scores, is available to assess the quality of student randomization for the initial experimental sample or for the subsequent additions to it. (The data in Figure 2D are equally consistent with either a true small class advantage or an initial assignment of somewhat better students to small kindergartens). It is also impossible to assess adequately the impact of differential attrition of experimental subjects, particularly of those in larger classes disappointed over their placement. Substantial, non-random test taking occurs over the years of the experiment. But, most important, the results depend fundamentally on the choice of teachers. While the teachers were to be randomly assigned to treatment groups, there is little description of how this was done. Nor is it easy to provide any reliable analysis of the teacher assignment, because only a few descriptors of teachers are found in the data and because there is little reason to believe that they adequately measure differences in teacher quality.[21] Moreover, teachers all knew they were participating in an experiment that could potentially affect the future resources available from the state. The schools themselves were self-selected and are clearly not random. Small schools were excluded from the study, and all participating schools were willing to provide their own partial funding to cover the full costs. (This school selection issue is important, because the STAR experiment heavily oversampled urban and minority schools where the achievement response to the program is thought to be largest).[22] The net result of each of these effects is difficult to ascertain, but there is prima facie evidence that the total impact is to overstate the impact of reduced class size (Hanushek 1999b).

The STAR experiment is very important from a methodological perspective, a point emphasized in Hanushek et al. (1994), Mosteller (1995), and Krueger (1999, 2000). More random-assignment experimentation is desperately needed in schools. But the evidence from this specific experiment should be interpreted with caution. Moreover, the evidence as it stands speaks just to the possible small effects of major and costly reductions in class size at kindergarten or first grade. It provides no evidence about beneficial effects at later grades. Nor does it indicate what effects could be expected from reductions of a smaller magnitude than the one-third reductions in Project STAR.

IV. Policy calculations

In addition to issues of how to interpret the existing class size evidence, Krueger (2000) attempts to provide a justification for undertaking large class size reductions even if the effects are as small as currently estimated by Project STAR. His argument is simple: small effects on achievement may have large enough impacts on subsequent earnings that the policies are justified. In order to do these calculations, Krueger takes the perspective that the proper comparison is between doing nothing and undertaking large reductions in class size. This perspective is very narrow and would lead to quite wasteful policies. Moreover, even to get to this justification, he must make a number of heroic assumptions about achievement and the labor market. These assumptions imply enormous uncertainty in the calculations, and thus in the subsequent policy recommendations.

Krueger (2000) presents a series of calculations based on chaining together a variety of uncertain estimates about key aspects of the rewards to higher achievement. In order to obtain estimates of the labor market returns to class size reductions, one must multiply the effect of the class size reduction on achieve-

ment times the impact of early achievement differences on performance throughout schooling and into the labor market. The subsequent estimates of initial labor market advantage must be projected across a person's working life and then discounted back to kindergarten to compare to the costs of the original class size reduction. The uncertainty with each of those steps grows when they are compounded together. The relationship between early achievement and subsequent earnings, for example, relies on a single study of British labor market experiences for a group of individuals born in 1958; their wages were recorded in 1981 and 1991.[23] These estimates are employed to project what expected early career labor market experiences might be in the United States around 2015, the relevant period for the policy deliberations. While it may be academically interesting to see if there is any plausibility to the kinds of class size policies being discussed, one would clearly not want to commit the billions of dollars implied by the policies on the basis of these back-of-the-envelope calculations.[24]

Surely improving achievement of students is very important and should be the focus of policy attention. The issue is not whether society should invest in quality but how it should invest. Calculations that suggest the economic justification is as close to breakeven as found by Krueger do not make a good case for the huge commitment of resources implicitly behind his calculations — particularly when the uncertainty of the calculations is recognized.

The heart of the issue, however, is that Krueger ignores the fact that existing evidence points to other factors — particularly teacher quality — as being more important than class size. The extensive research on student achievement over the past 35 years has made it clear that there are very important differences among teachers. This finding, of course, does not surprise many parents who are well aware of quality differences of teachers, but it has eluded many researchers. Researchers have tended to confuse measurability of specific teacher characteristics related to quality with real differences in quality. That is, the econometric research has not identified any teacher attributes (such as education, experience, background, type of training, certification, or the like) that are highly related to the ability of some teachers to get particularly large or particularly small gains in student learning. Nonetheless, econometric analyses have identified large and persistent differences in the effectiveness of different teachers.[25]

The magnitude of differences in teacher quality is impressive. For example, looking at the range of quality for teachers within a single large urban district, teachers near the top of the quality distribution can get an entire year's worth of additional learning out of their students compared to those near the bottom (Hanushek 1992).[26] That is, a good teacher will get a gain of one-and-a-half grade level equivalents, while a bad teacher will get a half year for a single academic year. A second set of estimates comes from recent work on students in Texas (Rivkin, Hanushek, and Kain 2000). This analysis follows several entire cohorts of students and permits multiple observations of different classes with a given teacher. We look at just the variations in student performance that arise from differences in teacher quality within a typical school and do not consider any variations across schools. The variation is large: moving from an average teacher to one at the 85th percentile of teacher quality (i.e., moving up one standard deviation in teacher quality) implies that the teacher's students would move up more than seven percentile rankings in the year.[27] These differences swamp any competing factors such as measured teacher and school attributes in their impact on student performance. For example, a one standard deviation reduction in class size implies a 0.01-.03 standard deviation improvement in student achievement. The lower-bound estimate on teacher quality summarized here implies that a one standard deviation change in quality leads to a 0.18 standard deviation increase in achievement. Finally, quality differences in teachers in Tennessee of a similar magnitude have also been estimated (Sanders and Horn 1995).

Recognizing the importance of teacher quality is central to the discussion of class size. First, any substantial reductions in class size imply hiring additional teachers. The success or failure of a class size reduction program will depend much more on whether or not the newly hired teachers are better or worse compared to the existing teachers than it will on the impact of class size reduction per se. In fact, depending upon the structure of the enabling legislation or policy, it could have quite detrimental effects.[28] Second, the Krueger calculations never consider the possibility of much more attractive alternatives to either the current schools or to class size reductions. Employing higher-quality teachers could produce major impacts on student performance that are unachievable with any realistic or feasible class size reductions.

A major difference in policies aimed at class size reduction and those aimed at changing teacher quality is their relationship to incentives in schools. There is ample reason to believe that the current incentives related to student performance are too weak (Hanushek et al. 1994). Essentially nobody within schools has much riding on whether or not students achieve at a high level. The expected pay and career of a good teacher is about the same as that for a bad teacher. Class size reduction does nothing to change this. On the other hand, if schools are to move toward attracting and retaining higher-quality teachers, they will almost certainly have to build in stronger performance incentives for school personnel. The exact form that this would take is unclear, and discussion of the options is beyond the scope of this paper (see, however, Hanushek et al. 1994). The necessity of altering incentives on the other hand seems clear, at least to economists.

Reducing class size does not logically preclude doing other things, but it is almost certainly a practical deterrent. Limited political attention and constraints on public funds imply that strong moves toward class size reduction are almost certain to drive out better policies aimed at improving teacher quality.

V. Conclusions

Despite the political popularity of overall class size reduction, the scientific support of such policies is weak to nonexistent. The existing evidence suggests that any effects of overall class size reduction policies will be small and very expensive. A number of investigations appear to show some effect of class size on achievement for specific groups or circumstances, but the estimated effects are invariably small and insufficient to support any broad reduction policies. The flawed analysis in Krueger (2000) does little to contribute to the debate on technical grounds and, more importantly, cannot change the inherent costs and expected benefits of the basic policy. The re-analysis of econometric estimates relies on placing heavy weight on lower-quality and biased econometric estimates. Even then, the efficacy of class size reduction is in doubt. The majority of his re-weighted estimates are still statistically insignificant, i.e., we have relatively little confidence that there is any effect on student outcomes. The most optimistic estimates suggest that the policy effects on student achievement would be small. The policy effects are shown by Krueger (2000) to make sense given the cost only if one makes a number of strong but uncertain assumptions and only if one believes that no other school policy is feasible.

Proposed class size reduction policies generally leave no room for localities to decide when and where reductions would be beneficial or detrimental. The existing evidence does not say that class size reductions are never worthwhile and that they should never be taken. It does say that uniform, across-the-board policies — such as those in the current policy debate — are unlikely to be effective.[29]

A significant problem is that there are few incentives that drive decisions toward ones that improve student performance. Most economists believe that incentives are key to results — whether in education or in other aspects of life. But schools are not organized in a way that they will decide to reduce class size in

instances where it is beneficial for student performance and not in other instances where it would not affect performance. Without such performance incentives, simply adding more resources is unlikely to lead to improvements in student achievement. In this regard, education has made very little progress in spite of the large and continuing investment in specific programs and activities.

Class size reduction is best thought of as a political decision. Past evidence suggests that it is a very effective mechanism for gaining voter support, even if past evidence also suggests that it is a very ineffective educational policy.

Appendix: Issues with the econometric data

Krueger (2000) raises a number of questions about the underlying estimates included in the overall summaries. Several of them were discussed with Krueger in private correspondence but did not make it into the published version.

Three coding questions are raised. First, as mentioned above, earlier correspondence determined that I had reversed the sign on the four estimated teacher-pupil ratio effects in Montmarquette and Mahseredjian (1989) in my previous tabulations, but Krueger subsequently does not make this correction in his tables. Second, Link and Mulligan (1986) included an ambiguous reference about whether teacher-pupil ratio was included in all 24 equations in their paper or just 12. Specifically, they noted that class size — which was discussed extensively in the modeling section — was insignificant in the mathematics equations, but they did not repeat mention of class size when they subsequently discussed the reading equations. In private communication with them designed to clarify this issue and to bring the most information to bear on the analysis, they indicated it was included in all 24 — and this was communicated to Krueger. Third, Kiesling (1967) is a journal article that extracted results from his thesis (Kiesling 1965), and the teacher-pupil ratio results came from his thesis. While this was noted in Hanushek (1986), it was not noted in Hanushek (1997), although it also was communicated to Krueger. (The omission of teacher-pupil ratio from the published article based on his thesis is a clear example of the publication bias discussed above. In this case it could be reliably avoided).

Endnotes

1. Pupil-teacher ratios are not the same as class size because of the use of specialist teachers, differences between numbers of classes taken by students and numbers taught by teachers, and other reasons. Nonetheless, because class size and pupil-teacher ratios tend to move together over time (see Lewit and Baker 1997) and because Krueger disregards any such distinctions, these differences are not highlighted at this time. See also Hanushek (1999a).

2. The NAEP has shown larger changes over time in the scores for 9- and 13-year-olds, but this has not been translated into improved scores at the end of high school; see Hanushek (1998) for further discussion.

3. Writing scores are first available in 1984. The mid-1980s saw a narrowing of the racial gap in achievement, but this stopped by 1990 and cannot be readily attributed to overall resource patterns. Further discussion of the aggregate trends including the racial trends can be found in Hanushek (1999a).

4. The analysis by Grissmer et al. (1994) attempts to aggregate these changes over time based on econometric estimates of how various family backgrounds affect achievement. This analysis indicates that the overall preparation of white students (based on family background factors) seems to have improved, while that for black students seems to have worsened. While considerable uncertainty surrounds the estimation approach, the analysis strongly suggests that changing backgrounds are not masking the effects of school resource increases. A critique of the methodology is found in Hanushek (1999a).

5. These tabulations were corrected for the previous miscoding of one article (Montmarquette and Mahseredjian 1989) that was pointed out to me by Alan Krueger. Krueger's analysis and tables of estimation results, however, do not adjust for this miscoding. A description of the criteria for inclusion is found in Hanushek (1997) and is summarized in Krueger (2000).

6. His discussion leads to some confusion in nomenclature. For reasons sketched below, my previous analyses have referred to distinct estimates as "studies" even though more than one estimate might appear in a given publication. Krueger changed this language by instead referring to separate publications as studies. Here I will generally drop the term studies and use the nomenclature of separate estimates in each publication.

7. In some of the published articles, an element of ambiguity about the exact estimation procedures and results exists. In tabulating sample sizes, for example, it was not clear whether the estimation in Harnisch (1987) was conducted at the individual student or the school level. Calculating its sample size on the basis of schools would increase the correlation between sample size and number of estimates in each publication to 0.10 and would provide a slightly different distribution of sample sizes in Table 3. While these changes are inconsequential for this discussion, more consequential ambiguities, such as those noted in Krueger (2000) and in the appendix, also exist. At times it was possible to resolve the ambiguities by bringing in outside information, which seemed to be the appropriate way to extract the most information from the existing publications.

8. This analysis also ignores statistically insignificant estimates for which the estimated sign is unknown, a condition making it impossible to know how to include them in the calculation. His analysis assumes that there is no information in analyses that drop further consideration of pupil-teacher ratios after an initial investigation.

9. This graph plots the Krueger results that do not correct the coding of Montmarquette and Mahseredjian (1989).

10. When important factors are omitted, estimates of the effect of varying teacher-pupil ratios will be unbiased only if there is no relationship across states between the quality of state policies and the average teacher-pupil ratio in the states. If on the other hand states with favorable education policies tend generally to have smaller classes, the estimates of teacher-pupil ratios will tend to differ systematically from the true effect of class size differences.

11. Hanushek, Rivkin, and Taylor (1996) demonstrate that any bias in the estimated parameters will be exacerbated by aggregation of the estimation sample. For example, 11 of the 277 estimates of the effects of teacher-pupil ratios come from highly aggregated performance and resource data measured at the state level, the level of measurement where policy information is omitted from the analyses.

12. Expenditure analyses virtually never direct analysis at performance across different classrooms or schools, since expenditure data are typically available only at the district level. Thus, they begin at a more aggregated level than many investigations of real resources.

13. In fact, using aggregate state data frequently precludes any consideration of different effects by student background, subject matter, or what have you — offering an explanation for why these publications have just one estimate.

14. While estimating some models with over a million observations, none is relevant for this analysis because each with a large sample fails to meet the eligibility criteria related to separating family background effects from correlated school resources (see below). In simple statistical terms, large samples cannot make up for estimating incorrectly specified relationships.

15. Other estimates rely on race to measure family background characteristics, but they consider the racial composition of observed schools or classrooms. Because parental education and income and other family attributes vary by race, including racial composition with measures of pupil-teacher ratios in these studies can begin to sort out causation from correlation in ways that Card and Krueger (1992b) cannot.

16. They also show that the results differ significantly across time and that they are very sensitive to the precise specification of the models. Speakman and Welch (1995) further show that virtually all of the effects of state school resources work through earnings of college attendees, even though the resource measures relate only to elementary and secondary schools.

17. Card and Krueger (1992a) is rightfully cited for its innovative combination of labor market data with school quality data. However, because it has been controversial, it is cited in other works (such as Heckman, Layne-Farrar, and Todd 1996a, 1996b) without providing any endorsement for its quality. A large number of citations are also of two different types. The first is its use in introductory material to justify a new set of estimates, as in: "while the common view is that resources do not matter, Card and Krueger find that they do." The second use is by other researchers who are looking to justify use of expenditure data in a different kind of analysis, say of school choice or school spending patterns. Neither is a statement about quality relative to other articles.

18. Surprisingly, policy discussions seldom focus on this finding about the ineffectiveness of teacher's aides.

19. Some students entered small classes in later grades, and their achievement was observed to be higher during their initial year of being in a small class than that of those in regular classes. See Hanushek (1999b) and Krueger (1999).

20. Throughout the four years of the experiment there was also substantial and nonrandom treatment group crossover (about 10% of the small class treatment group in grades 1-3). That is, some students originally assigned to large classes moved to small classes later in the experiment. A smaller number also went in the opposite direction. These students were clearly not random. While this problem can be dealt with analytically, it lowers the information that can be obtained from the experiment.

21. One measure of the importance of teachers relative to class size effects is that the average kindergarten achievement in small classes exceeds that in regular classes and regular-with-aide classes in only 40 of the 79 schools.

The teacher data include race, gender, teaching experience, highest degree, and position on the Tennessee career ladder. While there is no information about the effect of career ladder position on student performance, none of the other measures has been found to be reliable indicators of quality (Hanushek 1997). For estimates of the magnitude of variation in teacher quality, see below.

22. Krueger (1999) identifies significantly stronger effects for disadvantaged students, and these effects will then be overweighted in calculating program average treatment effects.

23. His discussion does consider two alternative estimates, although they appear to differ substantially from the estimates chosen for the calculations.

24. Krueger (2000) suggests that, because of uncertainty, it might be appropriate to compare his calculated rate of return to class size reductions to a somewhat higher interest rate than the 4% he appears to favor. His suggestion of perhaps considering a 6% return, however, vastly understates the uncertainty one would calculate by the normal procedure of developing confidence intervals for the estimates that enter into his illustrative benefit-cost approximations.

25. The econometric analysis behind these estimates involves calculating the average achievement gains across classrooms after allowing for differing student preparation, family background, and other factors. Some teachers consistently obtain high growth in student achievement, while others consistently obtain low growth. But standard measures of teacher characteristics are not correlated with quality measured in terms of value-added to student performance.

26. These estimates consider value-added models with family and school inputs. The sample includes only low-income minority students, whose average achievement in primary school is below the national average. The comparisons given compare teachers at the 5th percentile with those at the 95th percentile.

27. For a variety of reasons, these are lower-bound estimates of variations in teacher quality. Any variations in quality across schools would add to this. Moreover, the estimates rely on a series of conservative assumptions that all tend to lead to understatement of the systematic teacher differences.

28. The 1996 class size reduction program in California left inner city schools scrambling for new teachers, partly as a result of suburban districts' bidding away experienced teachers (Stecher and Bornstedt 1999). The likely net result is that disadvantaged students — the hypothesized winners from the reduction policy — actually suffered a loss in educational quality.

29. For example, the theoretical analysis of class size by Lazear (forthcoming) points to optimal policies when schools are trying to maximize student achievement. In this case, he shows that across-the-board reductions are never going to be the correct policy.

References

Burkhead, Jesse. 1967 *Input-Output in Large City High Schools.* Syracuse, N.Y.: Syracuse University Press.

Card, David, and Alan B. Krueger. 1992a. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100(1): 1-40.

Card, David, and Alan B. Krueger. 1992b. "School Quality and Black-White Relative Earnings: A Direct Assessment." *Quarterly Journal of Economics* 107(1): 151-200.

Card, David, and Alan B. Krueger. 1995. *Myth and Measurement: The New Economics of the Minimum Wage.* Princeton, N.J.: Princeton University Press.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of Educational Opportunity.* Washington, D.C.: U.S. Government Printing Office.

Congressional Budget Office. 1986. *Trends in Educational Achievement.* Washington, D.C.: Congressional Budget Office.

Finn, Jeremy D., and Charles M. Achilles. 1990. "Answers and Questions about Class Size: A Statewide Experiment." *American Educational Research Journal* 27(3): 557-77.

Grissmer, David W., Sheila Nataraj Kirby, Mark Berends, and Stephanie Williamson. 1994. *Student Achievement and the Changing American Family.* Santa Monica, Calif.: Rand Corporation.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141-77.

Hanushek, Eric A. 1992. "The Trade-Off Between Child Quantity and Quality." *Journal of Political Economy* 100(1): 84-117.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19(2): 141-64.

Hanushek, Eric A. 1998. "Conclusions and Controversies about the Effectiveness of School Resources." *FRBNY Economic Policy Review* 4 (March): 11-28.

Hanushek, Eric A. 1999a. "The Evidence on Class Size." In Susan E. Mayer and Paul Peterson, eds., *Earning and Learning: How Schools Matter.* Washington, D.C.: Brookings Institution.

Hanushek, Eric A. 1999b. "Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects." *Educational Evaluation and Policy Analysis* 21(2): 143-63.

Hanushek, Eric A., and Steven G. Rivkin. 1997. "Understanding the Twentieth-Century Growth in U.S. School Spending." *Journal of Human Resources* 32(1): 35-68.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor. 1996. "Aggregation and the Estimated Effects of School Resources." *Review of Economics and Statistics* 78(4): 611-27.

Hanushek, Eric A., et al. 1994. *Making Schools Work: Improving Performance and Controlling Costs.* Washington, D.C.: Brookings Institution.

Harnisch, Delwyn L. 1987. "Characteristics Associated With Effective Public High Schools." *Journal of Educational Research* 80(4): 233-41.

Heckman, James S., Anne Layne-Farrar, and Petra Todd. 1996a. "Does Measured School Quality Really Matter? An Examination of the Earnings-Quality Relationship." In Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* Washington, D.C.: Brookings Institution.

Heckman, James, Anne Layne-Farrar, and Petra Todd. 1996b. "Human Capital Pricing Equations With an Application to Estimating the Effect of Schooling Quality on Earnings." *Review of Economics and Statistics* 78(4): 562-610.

Hedges, Larry V. 1990. "Directions for Future Methodology." In Kenneth W. Wachter and Miron L. Straf, eds., *The Future of Meta-Analysis.* New York, N.Y.: Russell Sage.

Kiesling, Herbert. 1965. "Measuring a Local Government Service: A Study of School Districts in New York State." Ph.D. Dissertation, Harvard University, Cambridge, Mass.

Kiesling, Herbert. 1967. "Measuring a Local Government Service: A Study of School Districts in New York State." *Review of Economics and Statistics* 49 (August): 356-67.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B. 2000. "An Economist's View of Class Size Research." Mimeo.

Lazear, Edward. Forthcoming. "Educational Production." *Quarterly Journal of Economics.*

Lewit, Eugene M., and Linda Schuurmann Baker. 1997. "Class Size." *The Future of Children* 7(3): 112-21.

Link, Charles R., and James G. Mulligan. 1986. "The Merits of a Longer School Day." *Economics of Education Review* 5(4): 373-81.

Montmarquette, Claude, and Sophie Mahseredjian. 1989. "Does School Matter for Educational Achievement? A Two-Way Nested-Error Components Analysis." *Journal of Applied Econometrics* 4: 181-93.

Mosteller, Frederick. 1995. "The Tennessee Study of Class Size in the Early School Grades." *The Future of Children* 5(2): 113-27.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2000. "Teachers, Schools, and Academic Achievement." Working Paper No. 6691 (revised). Cambridge, Mass: National Bureau of Economic Research.

Sanders, William L., and Sandra P. Horn. 1995. "The Tennessee Value-Added Assessment System (TVAA): Mixed Model Methodology in Educational Assessement." In Anthony J. Shinkfield and Daniel L. Stufflebeam, eds., *Teacher Evaluation: Guide to Effective Practice.* Boston, Mass.: Kluwer Academic Publishers.

Speakman, Robert, and Finis Welch. 1995. "Does School Quality Matter? A Reassessment." Texas A&M University. Mimeo.

Stecher, Brian M., and George W. Bohrnstedt, eds. 1999. *Class Size Reduction in California: Early Evaluation Findings, 1996-98.* Palo Alto, Calif: American Institutes for Research.

Word, Elizabeth, John Johnston, Helen Pate Bain, B. DeWayne Fulton, Jayne Boyd Zaharies, Martha Nannette Lintz, Charles M. Achilles, John Folger, and Carolyn Breda. 1990. *Student/Teacher Achievement Ratio (STAR), Tennessee's K-3 Class Size Study: Final Summary Report, 1985-1990.* Nashville: Tennessee State Department of Education.